


ARTICLE

<https://doi.org/10.1038/s41467-020-14482-y>

OPEN

# Determining sequencing depth in a single-cell RNA-seq experiment

Martin Jinye Zhang <sup>1,3</sup>, Vasilis Ntranos<sup>1,2,3</sup> & David Tse<sup>1\*</sup>

An underlying question for virtually all single-cell RNA sequencing experiments is how to allocate the limited sequencing budget: deep sequencing of a few cells or shallow sequencing of many cells? Here we present a mathematical framework which reveals that, for estimating many important gene properties, the optimal allocation is to sequence at a depth of around one read per cell per gene. Interestingly, the corresponding optimal estimator is not the widely-used plug-in estimator, but one developed via empirical Bayes.

---

<sup>1</sup>Department of Electrical Engineering, Stanford University, Stanford, CA, USA. <sup>2</sup>Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA. <sup>3</sup>These authors contributed equally: Martin Jinye Zhang, Vasilis Ntranos. \*email: [dntse@stanford.edu](mailto:dntse@stanford.edu)

Single-cell RNA sequencing (scRNA-seq) technologies have revolutionized biological research over the past few years by providing the tools to simultaneously interrogate the transcriptional states of thousands of cells in a single experiment. In contrast to bulk RNA-Seq, which probes the average gene expression in a cell population, single-cell RNA-seq has unlocked the potential of extracting higher-order information, granting us access to the underlying gene expression distribution. Indeed, this unprecedented look into population-level heterogeneity has been vital in the success of scRNA-seq, leading up to new biological discoveries<sup>1,2</sup>.

Although early single-cell RNA-seq assays were labor intensive and initially constrained by the small number of cells that could be processed in a single experiment, recent technological advances have allowed hundreds of thousands of cells to be assayed in parallel<sup>3</sup>, eliminating the otherwise prohibitive per cell cost overhead. From a sequencing budget perspective, however, this seemingly unconstrained increase in the number of cells available for scRNA-seq introduces a practical limitation in the total number of reads that can be sequenced per cell. More reads can significantly reduce the effect of the technical noise in estimating the true transcriptional state of a given cell, whereas more cells can provide us with a broader view of the biological variability in the cell population. A natural experimental design question arises (Fig. 1a): how many cells should we choose to profile for a given study, and at what sequencing depth?

The experimental design question has attracted a lot of attention in the literature<sup>4–8</sup>, but as of now, there has not been a clear answer. Several studies provide evidence that a relatively shallow sequencing depth is sufficient for common tasks such as cell type identification and principal component analysis (PCA)<sup>9–11</sup>, whereas others recommend deeper sequencing for accurate gene expression estimation<sup>12–15</sup>. Despite the different recommendations, the approach to providing experimental design guidelines is shared among all: given a deeply sequenced dataset with a pre-defined number of cells, how much subsampling can a given method tolerate? An example of this conventional approach is also evident in the mathematical model used in a recent work<sup>11</sup> to study the effect of sequencing depth on PCA. Although practically relevant, this line of work does not provide a comprehensive solution to the underlying experimental design question because of three reasons: (1) the number of cells is fixed and implicitly assumed to be enough for the biological question at hand; (2) the deeply sequenced dataset is considered to be the ground truth; (3) the corresponding estimation method is chosen a priori and is tied to the experiment.

In this work, we propose a mathematical framework for single-cell RNA-seq that fixes not the number of cells but the total sequencing budget, and disentangles the biological ground truth from both the sequencing experiment as well as the method used to estimate it. In particular, we consider the output of the sequencing experiment as a noisy measurement of the true underlying gene expression and evaluate our fundamental ability to recover the gene expression distribution using the optimal estimator. The two design parameters in our proposed framework are the total number of cells to be sequenced  $n_{\text{cells}}$  and the sequencing depth in terms of the total number of reads per cell  $n_{\text{reads}}$ , both affecting the optimal estimation error. Now, the experimental design tradeoff becomes apparent when these two quantities are tied together under a total sequencing budget constraint  $B = n_{\text{cells}} \times n_{\text{reads}}$  (Fig. 1a, sequencing budget allocation problem). The sequencing budget  $B$  corresponds to the total number of reads that will be generated and is directly proportional to the sequencing cost of the experiment (see Methods).

More specifically, we consider a hierarchical model<sup>16–18</sup> to analyze the tradeoff in the sequencing budget allocation problem

(see Methods). At a high level, we assume an underlying high-dimensional gene expression distribution  $P_{\mathbf{X}}$  that carries the biological information of the cell population we are interested in and is independent of the sequencing process (Fig. 1a top). The cells  $1, 2, \dots$  in the experiment are described by gene expressions  $\mathbf{X}_1, \mathbf{X}_2, \dots$  sampled from  $P_{\mathbf{X}}$ , whereas we can only observe the read counts  $\mathbf{Y}_1, \mathbf{Y}_2, \dots$  that are generated from the corresponding gene expressions via sequencing (Fig. 1a bottom). In this context, it is clear that with many cells  $n_{\text{cells}}$  we can estimate the read count distribution  $P_{\mathbf{Y}}$  accurately, whereas with more reads per cell  $n_{\text{reads}}$  we can make sure that the individual (normalized) observations  $\mathbf{Y}_1/n_{\text{reads},1}, \mathbf{Y}_2/n_{\text{reads},2}, \dots$  are much closer to the ground truth expressions  $\mathbf{X}_1, \mathbf{X}_2, \dots$  of the cells (here,  $n_{\text{reads},c}$  represents the total number of reads for cell  $c$  and the average of  $n_{\text{reads},c}$  over all cells is  $n_{\text{reads}}$ ). The optimal tradeoff is then derived to reconcile the two.

## Results

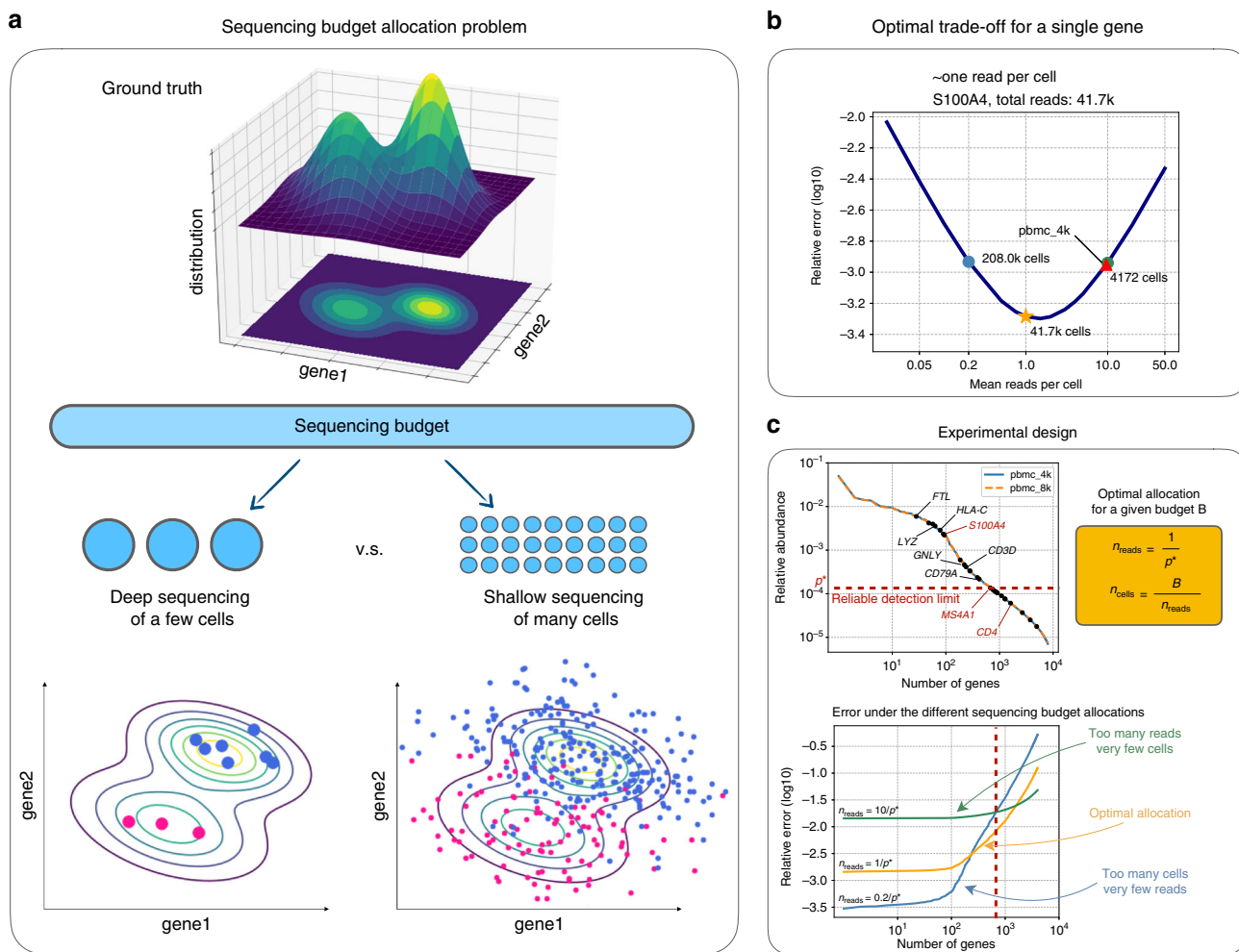
**Model overview.** The gene expression levels of each cell, denoted by  $\mathbf{X}_c = [X_{c1}, \dots, X_{cG}]$  for  $c = 1, \dots, n_{\text{cells}}$ , can be viewed as independent samples from the gene expression distribution  $P_{\mathbf{X}}$ , where  $G$  denotes the number of genes. More specifically, we assume that  $X_{cg}$  represents the true relative abundance of the mRNA molecules originating from a gene  $g$  in cell  $c$ , so that  $\sum_{g=1}^G X_{cg} = 1$ . To model the sequencing process, we assume that after a particular cell  $\mathbf{X}_c$  has been sampled from  $P_{\mathbf{X}}$ , its corresponding gene counts  $\mathbf{Y}_c = [Y_{c1}, \dots, Y_{cG}]$  are generated via Poisson sampling of  $\gamma_c \cdot n_{\text{reads}}$  reads from  $\mathbf{X}_c$ , where  $\gamma_c$  is a size factor that is cell-specific but not gene-specific. Overall, our hierarchical model is given by (here, we simplified the model by fixing  $\gamma_c$ ; see Eq. (2) in the Methods section for the complete model and Supplementary Note 1 for more details): for cells  $c = 1, 2, \dots, n_{\text{cells}}$ ,

$$\mathbf{X}_c \sim P_{\mathbf{X}}, \quad \text{and} \quad Y_{cg} | X_{cg} \sim \text{Poi}(\gamma_c n_{\text{reads}} X_{cg}) \text{ for } g = 1, 2, \dots, G. \quad (1)$$

Under this framework, the ultimate goal of a single-cell RNA-seq experiment would be to estimate quantities related to the (unknown) ground truth distribution  $P_{\mathbf{X}}$  from the noisy measurements  $\mathbf{Y}_c$ . Fixing the total sequencing budget  $B = n_{\text{cells}} \times n_{\text{reads}}$ , we aim to characterize the optimal experimental design tradeoff between the number of cells  $n_{\text{cells}}$  and the number of reads per cell  $n_{\text{reads}}$  that can minimize the corresponding estimation error.

Although our framework is non-parametric—in the sense that no particular prior is assumed for the underlying gene distribution  $P_{\mathbf{X}}$ , it is instructive to illustrate the framework in the context of the widely used overdispersion model, where for each gene  $g$ , the read counts  $Y_{cg}$  are assumed to follow a negative binomial distribution<sup>19–21</sup>. As the negative binomial distribution can be derived as a gamma–Poisson mixture, the resulting model can be viewed as a special case of Eq. (1) in which the underlying gene expression marginals follow gamma distributions. In that case, one would be interested in estimating the marginals  $X_g \sim \text{Gamma}(r_g, \theta_g)$ , effectively decoupling the true biological variability from the technical noise that was introduced during sequencing via Poisson sampling (see Relation to the overdispersion model in Supplementary Note 4).

As a technical remark, for assuming the gamma–Poisson mixture, we dropped two constraints without loss of generality, i.e.,  $X_{cg} < 1$  and  $\sum_{g=1}^G X_{cg} = 1$ . The former is because the relative expression  $X_{cg}$  is of the order of  $1/G$ , which is much smaller than 1. With a mean much smaller than 1, the truncated gamma distribution with truncation at 1 is very close to



**Fig. 1 Optimal sequencing budget allocation.** **a** Description of the sequencing budget allocation problem. Consider estimating the underlying gene distribution (top) from the noisy read counts obtained via sequencing (bottom). With a fixed number of reads to be sequenced, deep sequencing of a few cells accurately estimates each individual cell but lacks coverage of the entire distribution (left), whereas a shallow sequencing of many cells covers the entire population but introduces a lot of noise (right). **b** Optimal tradeoff. The memory T-cell marker gene *S100A4* has 41.7k reads in the pbmc\_4k dataset. For estimating the underlying gamma distribution  $X_g \sim \text{Gamma}(r_g, \theta_g)$ , the relative error is plotted as a function of the sequencing depth, where the optimal error is obtained at a depth of one read per cell (orange star) and is two times smaller than that at the current depth of pbmc\_4k (red triangle). **c** Experimental design. To determine the sequencing depth for an experiment, first the relative gene expression level can be obtained via pilot experiments or previous studies (top left). Then the researcher can select a set of genes of interest (i.e., some marker genes highlighted as black dots), of which the smallest relative expression level  $p^*$  (*MS4A1*) defines the reliable detection limit. Finally, the optimal sequencing depth is determined as  $n_{reads}^* = 1/p^*$  (top right). The errors under different tradeoffs are visualized as a function of the genes ordered from the most expressed to the least (bottom). The optimal sequencing budget allocation (orange) minimizes the worst-case error over all the genes of interest (left of the red dashed line), whereas both the deeper sequencing (green) and the shallower sequencing (blue) yield worse results.

the corresponding untruncated distribution. The latter is because that the number of genes  $G$  is large, and therefore,  $\sum_{g=1}^G X_{cg}$  concentrates around its mean, which is 1.

**Optimal sequencing budget allocation.** For our main results, we focused on 3'-end sequencing technologies<sup>22–24</sup> and used the above framework to study the experimental design tradeoff for estimating several important quantities of the underlying gene distribution, such as the CV and the Pearson correlation (see the Methods section). In the context of 3'-end sequencing,  $P_X$  naturally models the unknown high-dimensional distribution of mRNA abundances across cells, whereas the read counts for the cells,  $Y_1, Y_2, \dots$ , correspond to the number of unique molecular identifiers (UMIs) observed via sequencing. Our main result states that the optimal budget allocation (i.e., the one that

minimizes the estimation error) is achieved by maximizing the number of cells while making sure that at least  $\sim 1$  UMI per cell will be observed on average for all genes of primary biological interest in the experiment.

As a demonstrating example, in Fig. 1b we consider the memory T-cell marker gene *S100A4* to be of primary biological interest and evaluate the optimal tradeoff in the context of the overdispersion model for the total sequencing budget used to generate the 10x Genomics' pbmc\_4k dataset (4340 cells, total 41.7 k reads for *S100A4*); our analysis suggests that the optimal tradeoff would have been attained by sequencing 10 times shallower using 10 times more cells, reducing the error by twofolds. Of course, the recommended sequencing depth depends on the genes under consideration. For example, the sequencing depth of pbmc\_4k dataset is optimal when the B-cell marker gene *MS4A1* is considered, and it should be sequenced four times

deeper with 1/4 cells when the T-helper marker gene *CD4* is considered (Fig. 1c top, Supplementary Fig. 2a, b). The latter arguably has reached saturation for the 10x Genomics' technology. Hence, the guidance there is to sequence until saturation, i.e., sequence until no more new UMIs are observed (see Experimental design in the Methods section as well as Supplementary Note 3).

As the example indicates, an important aspect of our framework is to allow flexible experimental design at a single-gene resolution. The researcher can thus design the experiment based on the mean expression level of a set of important genes related to the biological question, where the mean expression level can be obtained via pilot experiments or previous studies (see Experimental design in the Methods section). We illustrate the proposed experimental design procedure by considering peripheral blood mononuclear cells (PBMCs) with the corresponding marker genes (Fig. 1c). The goal is to ensure reliable estimation for all these genes that are above a certain expression level, say that of *MS4A1*. Hence, the expression level of the gene of interest, i.e., *MS4A1*, naturally defines the reliable detection limit  $p^*$  at which we should guarantee observation of one average UMI per cell. Thus, given a budget  $B$ , choosing  $n_{\text{reads}}^* = 1/p^*$  and  $n_{\text{cells}}^* = B/n_{\text{reads}}^*$  achieves the optimal tradeoff for reliable detection at  $p^*$ . In this example, *MS4A1* will be sequenced  $\sim 1$  UMI per cell on average. Interestingly, this approach suggests a slightly deeper sequencing for current 10x datasets (Supplementary Figs. 1 and 2).

Unlike estimating the gamma distribution parameters for the overdispersion model, we considered estimating other quantities in a non-parametric setting (see also a non-parametric interpretation of the overdispersion model in Relation to the overdispersion model in Supplementary Note 4). Although the exact optimal depth is task-dependent, our empirical evaluations have shown that the above recommendation is remarkably consistent across all quantities considered in this paper—typically lying in a narrow range between 0.2 and 1 (Fig. 2a, Supplementary Fig. 4). Last but not the least, our tradeoff analysis can also provide a post hoc guidance for reliable estimation for existing datasets, namely for certain quantities, to determine which genes can be reliably estimated and which cannot, based on their mean expression level (Fig. 2b, see also post hoc guidance for reliable estimation in Methods).

**Optimal estimator.** Another important result arising from our experimental design framework is the fundamental role of the estimator in the optimal tradeoff. A very common—almost routine—practice in the literature is to use the so-called plug-in estimator, which, as a general recipe, blindly uses the scaled (relative) read counts  $\mathbf{Y}_1/n_{\text{reads},1}, \mathbf{Y}_2/n_{\text{reads},2}, \dots$  as a proxy for the true relative gene expression levels  $\mathbf{X}_1, \mathbf{X}_2, \dots$ , effectively estimating the corresponding distributional quantities by plugging-in the observed values. For example, the plug-in estimator naturally estimates the mean of the gene expression distribution  $P_X$  by that of  $P_{\mathbf{Y}/n_{\text{reads}}}$ , the variance of  $P_X$  by that of  $P_{\mathbf{Y}/n_{\text{reads}}}$ , etc. This approach, although very accurate for deeply sequenced datasets, becomes increasingly problematic in the limit of shallow sequencing; overdispersion and inflated dropout levels in lowly expressed genes, typically associated in the literature with scRNA-seq, are some of the more pronounced consequences.

For the sequencing budget allocation problem, we did not restrict our results to any particular estimator; our analysis suggested that the optimal tradeoff cannot be achieved by the conventional plug-in approach but with another class of estimators developed via empirical Bayes modeling<sup>16–18,25,26</sup> (see Methods). Such estimators are inherently aware of the

Poisson sampling noise introduced by sequencing, and therefore can adapt to various sequencing depths. As they estimate the prior gene distribution  $P_X$  in the hierarchical model (2) from the observed data  $\mathbf{Y}_c$ , sometimes by estimating the moments of the prior distribution  $P_X$ , they are usually associated with the names empirical Bayes, moment matching, or density deconvolution. Here, we use the term EB to refer to them in general.

In Figs. 3 and 4 (also Supplementary Figs. 7–12), we provide a comprehensive evaluation of the performance of EB estimators in several key applications and show that they provide remarkably consistent estimates across varying sequencing depths and different datasets. Also, they are shown to be biologically meaningful (Fig. 4c, Supplementary Fig. 12). In contrast, the plug-in approach, being sensitive to the sequencing depth, significantly overestimates the variability in gene expression (CV) owing to the inevitable zero-inflation occurring at shallow sequencing (Fig. 3a), and subsequently limits the performance of common downstream tasks such as PCA and gene network analysis (Methods, Fig. 3b, Fig. 4).

**Validation against the gold standard smFISH.** In order to further validate our results that the optimal sequencing depth is attained at  $\sim 1$  average UMI per cell, and that the EB estimates are indeed close to the ground truth, we considered two additional datasets<sup>15,27</sup>, accompanied by the single-molecule fluorescent in situ hybridization (smFISH) data, which is regarded as the gold standard for measuring the number of mRNAs in a cell. The libraries for the two scRNA-seq datasets were generated by Drop-seq<sup>23</sup> and CEL-seq<sup>28</sup>, respectively, two UMI-based technologies.

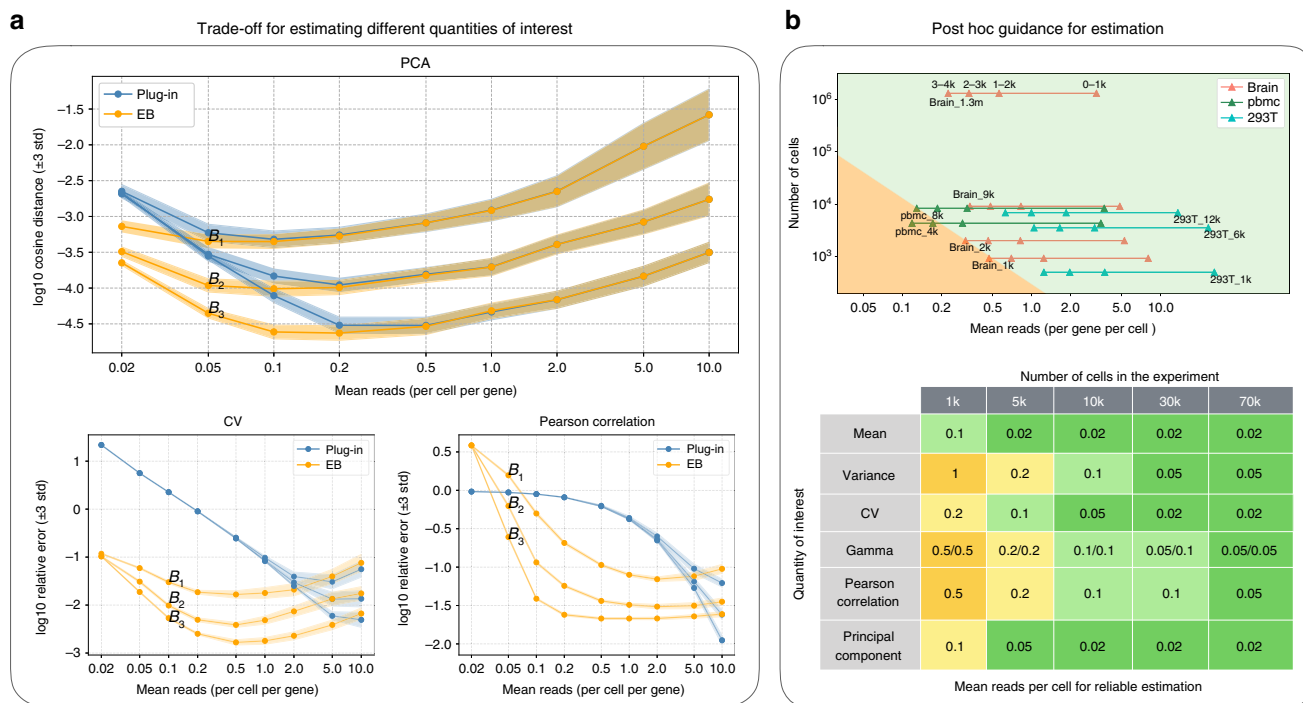
We first compared the estimated CV and inactive probability against the smFISH estimates. The EB estimates agree well with the smFISH data while there is clear inflation for the plug-in estimates (Fig. 5a, Supplementary Fig. 13). Furthermore, we investigated the optimal sequencing depth by fixing the budget  $B$  and varying the sequencing depth, as we did in Fig. 2a. The critical difference here is that the error is evaluated against the smFISH data, which serves as a proxy for the ground truth. Two genes, *MITF* and *VGF*, were considered that have relatively more UMIs to subsample. The tradeoff curves (Fig. 5b, Supplementary Fig. 14) are qualitatively similar to the simulation studies (Fig. 2a, Supplementary Fig. 4), showing an optimal depth between 0.1 and 0.6. This is consistent with the experimental design guidelines that we provided in our earlier analysis.

See also Supplementary Figs. 15–17 and the Details of the ERCC experiments subsection in Supplementary Note 6 for additional validations using datasets with ERCC synthetic spike-in RNAs and pure RNA controls.

## Discussion

A natural yet challenging experimental design question for single-cell RNA-seq is how many cells should one choose to profile and at what sequencing depth to extract the maximum amount of information from the experiment. In this paper, we introduced the sequencing budget allocation problem to provide a precise answer to this question; given a fixed budget, sequencing as many cells as possible at approximately one read per cell per gene is optimal, both theoretically and experimentally.

Conceptually, there are three important aspects of our mathematical framework that enabled our theoretical analysis and led to the development of the corresponding sequencing-depth-aware EB estimators. First, we explicitly incorporated the notion of an unknown ground truth distribution  $P_X$  that describes the underlying single-cell population of interest. From this perspective, a single-cell RNA-seq experiment can be naturally seen as an attempt to extract information about this distribution. Second, we



**Fig. 2 Empirical quantification of the optimal sequencing depth.** **a** Simulations of error under different budget allocation. 3-std confidence intervals are provided. The top panel simulates the error for estimating the first principal direction using the plug-in estimator (blue) and the EB estimator (orange), respectively. Three budgets are considered, i.e.,  $B_1 = 0.6$  k per gene,  $B_2 = 3$  k per gene,  $B_3 = 15$  k per gene. The depth (mean reads per cell per gene) ranges from 0.02 to 10. The result indicates that the optimal depth for the EB estimator is the same (-0.1) for all three budgets, validating the theory that the optimal depth is independent of the budget. The cases for the coefficient of variation and the Pearson correlation (bottom) also show similar qualitative behaviors. **b** Post hoc guidance for reliable estimation. We visualized the top 4k genes of some representative datasets (top), where a triangle residing in the green region means the Pearson correlation of corresponding genes can be reliably estimated (relative error < 10%). For example, we can reliably estimate the first 2k genes for the brain\_1k dataset and all 4k genes for the brain\_9k dataset. A more comprehensive result is summarized in the bottom table. For example, the first element (mean, 1k) shows that with 1k cells, a gene needs to have at least 0.1 reads per cell for reliably estimating the mean.

disentangled this biological ground truth not only from the sequencing process but also from the method used to estimate it. Considering the output of the sequencing experiment as a noisy measurement  $P_Y$  of the true underlying distribution, we were able to mathematically evaluate our fundamental ability to recover  $P_X$  and identify the corresponding tradeoff-optimal estimators for several quantities of interest by essentially optimizing over all possible methods and experimental design parameters. Finally, to provide practical experimental design guidelines, we considered how different biological questions could be incorporated within our framework. Assuming that a biological question can be defined in terms of a set of genes of interest (e.g., associated with a particular pathway), we were able to provide sequencing depth recommendations by minimizing the worst-case error within that set.

Our experimental results showed that the proposed EB estimators could achieve significantly better performance compared with the conventional plug-in approach that is commonly used by existing single-cell analysis methods. Importantly, we demonstrated that the proposed estimators produce unbiased results across deep and shallow datasets obtained from the same underlying population of cells and validated their ability to produce estimates that are very close to the ground truth as measured by smFISH. We also provided post hoc guidance for reliable estimation by evaluating our results on multiple genes from different biological samples. Apart from providing cost-efficient data generation guidelines for future experiments, we believe that our results are also going to be useful in assessing the quality and statistical interpretability of existing datasets, particularly in the context of global collaborative initiatives such as the Human Cell Atlas<sup>29</sup>.

**Methods**

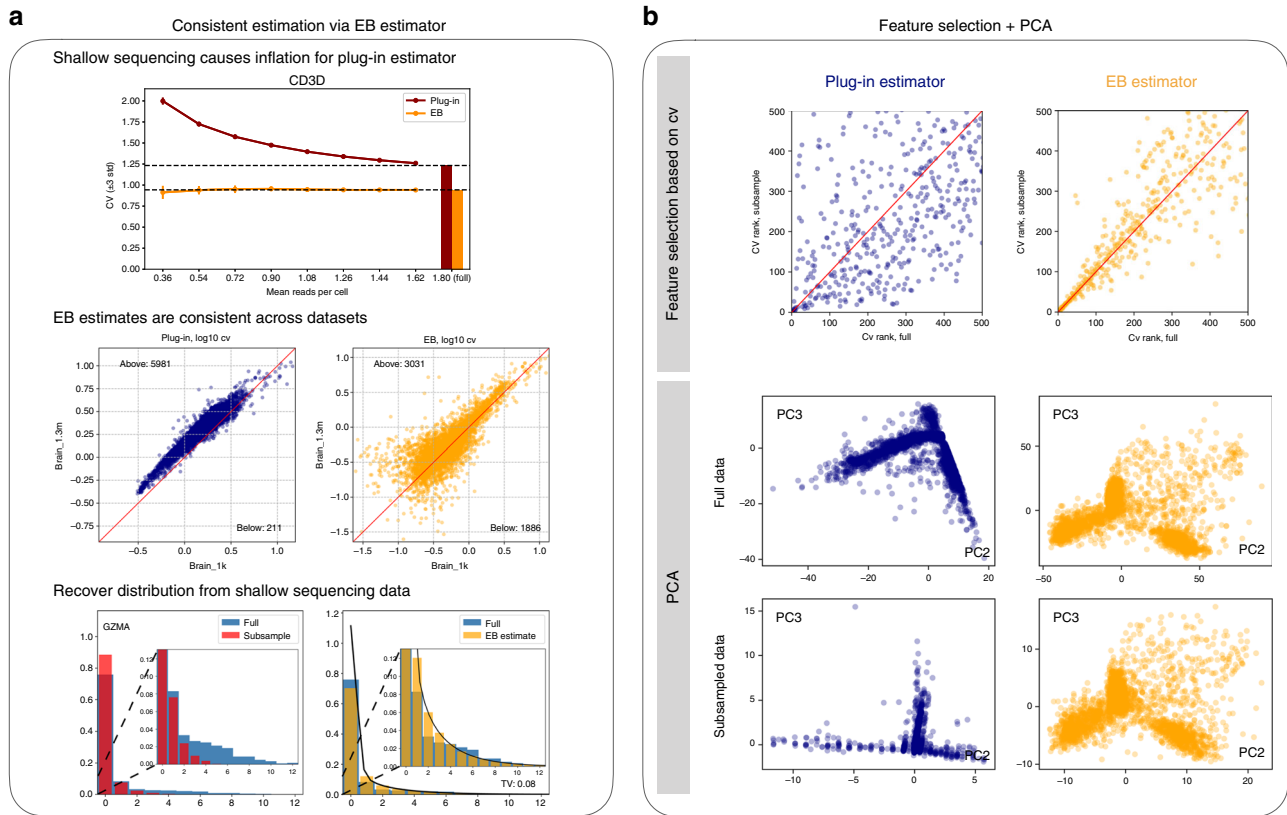
**Model.** For a scRNA-seq experiment, let  $n_{\text{cells}}$  be the number of cells and  $n_{\text{reads}}$  be the average UMIs per cell. The total number of UMIs  $B = n_{\text{cells}} \times n_{\text{reads}}$  is used to denote the available budget for this experiment. Given a fixed budget, we are interested in the optimal allocation between  $n_{\text{cells}}$  and  $n_{\text{reads}}$  for estimating certain distributional quantities that are important to the scRNA-seq analysis.

We adopt a hierarchical model for the analysis. Let  $G$  be the number of genes and for each cell  $c = 1, \dots, n_{\text{cells}}$ , let  $X_c = [X_{c1}, \dots, X_{cG}]$  be the relative gene expression level satisfying  $\sum_{g=1}^G X_{cg} = 1$ . The relative expression levels are assumed to be drawn i.i.d. from some unknown cell distribution  $P_X$ , which is defined with respect to the cell population under investigation—it may be cells from a certain tissue or some isolated cell sub-populations. This is quite a general model. For example, cells coming from several sub-populations can be modeled by letting  $P_X$  be a mixture distribution. The gene expression level  $X_c$  is measured by the observed UMIs  $Y_c \in \mathbb{N}^G$  via sequencing, of which the stochastic process is modeled using Poisson noise; such noise model has been extensively validated by previous works<sup>16,30</sup>. In addition, we assume a size factor  $\gamma_c$  for each cell that accounts for the variation in cell sizes. To summarize, for gene  $g = 1, \dots, G$  in cell  $c = 1, \dots, n_{\text{cells}}$ , we have assumed

$$X_c \stackrel{\text{i.i.d.}}{\sim} P_X, \gamma_c \stackrel{\text{i.i.d.}}{\sim} P_\gamma, Y_{cg} | X_{cg}, \gamma_c \sim \text{Poi}(\gamma_c n_{\text{reads}} X_{cg}). \tag{2}$$

**Quantities to estimate.** We study the optimal sequencing budget allocation for estimating the following distributional quantities of  $P_X$  that are commonly used in scRNA-seq analysis. See Supplementary Note 2 for more details.

- The marginal gene moments  $M_{k,g} = \mathbb{E}[X_{cg}^k], g = 1, \dots, G, k = 1, 2, \dots$ . The marginal gene moments can be used to compute quantities like the mean expression, CV, the Fano factor, or the parameters for the overdispersion model (assuming that  $X_{cg}$  follows a gamma distribution), which play an important role in data pre-processing, feature selection, and gene-type identification<sup>31,32</sup>.
- The gene expression covariance matrix  $K \in \mathbb{R}^{G \times G}$ , which also gives rise to the Pearson correlation matrix. Both quantities can be used to study the dependency structure of genes, e.g., via spectrum methods<sup>33–35</sup> or gene network analysis<sup>36,37</sup>.



**Fig. 3 EB estimates are consistent between deep and shallow datasets.** **a** Top: for estimating the coefficient of variation (CV), the plug-in estimates become more inflated as the sequencing depth becomes shallower (from right to left along the x axis), whereas the EB estimates are consistent. 3-st confidence intervals are provided for this panel. Middle: both brain\_1k and brain\_1.3m are from the mouse brain, and hence each gene should have a similar CV value between the two datasets. This is indeed the case for the EB estimator (right), which is adaptive to different sequencing depths. However, as brain\_1k is twice deeper than brain\_1.3m, the plug-in estimates are biased that most points are above the 45-degree line (red). Bottom: distribution recovery for the gene GZMA from a dataset that is subsampled to be five times shallower (left). The EB estimator provides a reasonable estimation for both the zero proportion and the tail shape, resulting in a small total variation error (right). **b** Feature selection and PCA. The task is to first select features (genes) based on CV, and then perform PCA on the selected features. The results on the full data (pbmc\_4k) and a subsampled (three times shallower) are compared. EB estimates are more consistent between the full data and the subsampled data for both the CV ranks (top) and the PCA plots (bottom).

- The inactive probability of a gene  $g$  with the definition  $p_{0,g}(\kappa) = \mathbb{E}[\exp(-\kappa X_{cg})]$ . It also has the interpretation of the proportion of zero-UMI cells for gene  $g$  when the cell population is sequenced  $\kappa/n_{reads}$  times deeper. As special cases,  $\kappa = \infty$  corresponds to the probability that  $X_{cg}$  is zero, whereas  $\kappa = n_{reads}$  corresponds to the proportion of cells whose observed counts  $Y_{cg}$  are zero. The latter was also considered in a recent work<sup>38</sup>.
- The inactive probability of a gene pair  $g_1, g_2$  with the definition  $p_{0,g_1,g_2}(\kappa) = \mathbb{E}[\exp(-\kappa(X_{cg_1} + X_{cg_2}))]$  that quantifies the change that both genes are inactive. This quantity can be used to analyze the gene co-expression network<sup>38</sup>.
- The marginal gene distribution  $P_{X_g}$  (also considered in a recent work<sup>16</sup>).

**Optimal sequencing budget allocation.** We considered a single gene and derived the optimal budget allocation for estimating all the above quantities of its distribution  $P_{X_g}$  (see Supplementary Note 5 for more details). As the mean relative expression level of a gene  $g$  is relatively stable within a specific tissue/sample (see Experimental design subsection below for more details), one can safely estimate that for an experiment with budget  $B$ , the total number of reads for gene  $g$  is around  $p_g B$ . Then the tradeoff with respect to gene  $g$  can be written as  $p_g B = n_{reads,g} \times n_{cells}$ , where  $n_{reads,g}$  is the mean reads per cell for gene  $g$ , satisfying the relation  $n_{reads,g}^* = p_g n_{reads}$ .

**Theorem 1.** (Optimal budget allocation, informal) For estimating moments, covariance matrix, inactive probability, pairwise inactive probability, and distribution, the optimal budget allocation is

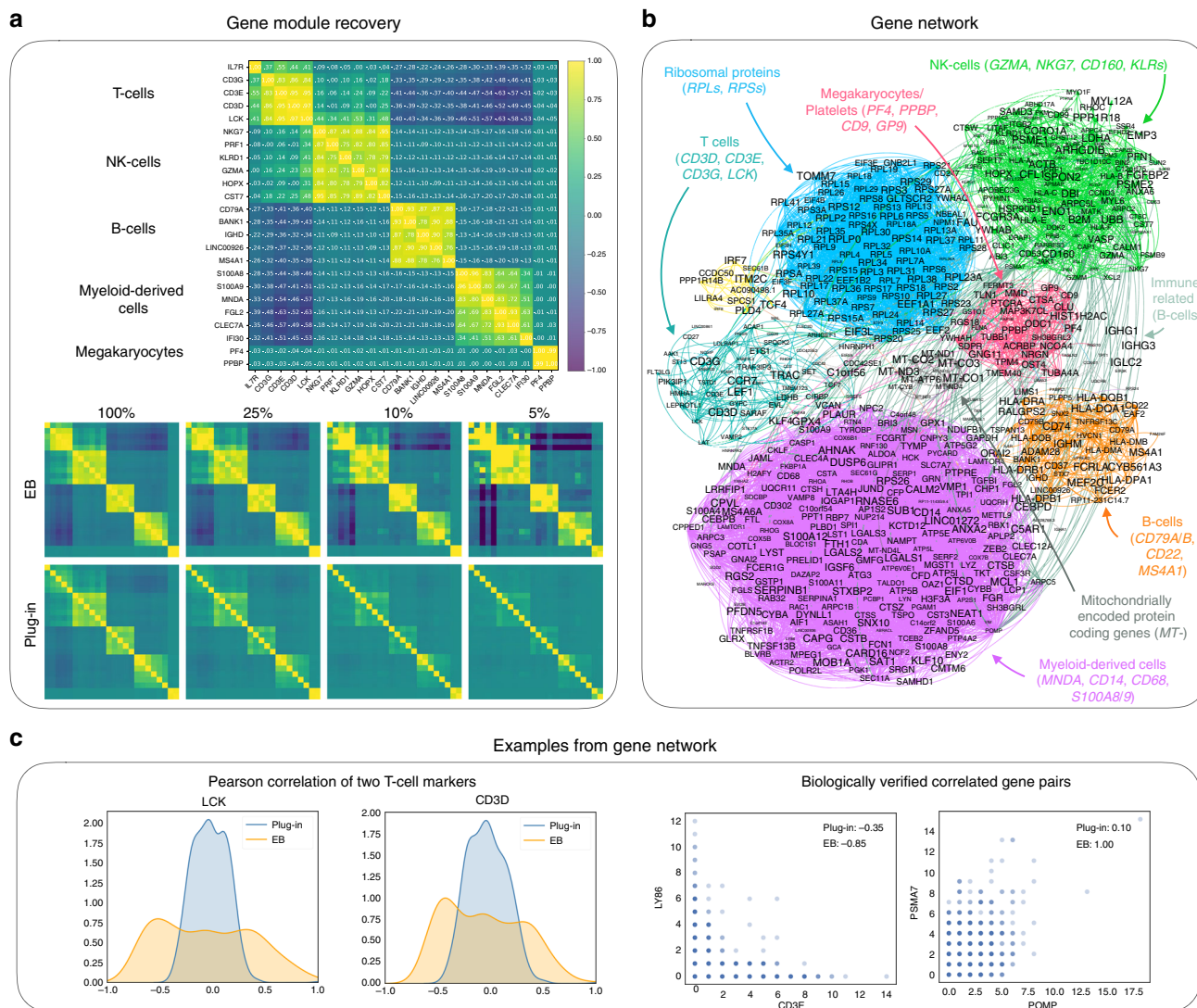
$$n_{reads,g}^* \sim 1, n_{cells}^* \sim B p_g. \quad (3)$$

The optimality is in the sense of minimizing the worst-case error over a family of distributions  $P_{X_g}$  with mild assumptions and the optimal error rate is achieved by the EB estimators.

The expression  $n_{reads,g}^* \sim 1$  in Theorem 1 implies that the optimal sequencing depth (mean reads per cell per gene) is given by some constant independent of the sequencing budget (see the formal statement in Supplementary Note 5). Therefore, for a scRNA-seq experiment, we should aim at a certain sequencing depth; when the budget increases, we should keep the same depth and allocate the additional budget toward collecting more cells. In other words, after having achieved a certain sequencing depth, deeper sequencing does not help as much as having more cells. We also note that the actual value of this optimal sequencing depth may be different for estimating different quantities, which is further investigated in the following section. In addition, Theorem 1 suggests that the EB estimators should be used for optimal estimation, whose effectiveness is demonstrated in Figs. 3–4.

**Experimental design.** The exact values of the optimal sequencing depth  $n_{reads,g}^*$  for estimating different quantities were investigated both theoretically and via simulations. First, the closed-form expressions of the optimal depth  $n_{reads,g}^*$  were derived for estimating the mean, the second moment, and the gamma parameters (of overdispersion model), which depend on the distribution  $P_{X_g}$  but are nonetheless  $\sim 1$  for typical cases (Supplementary Notes 3 and 5). Second, estimation errors under different budget splits were simulated by subsampling from a real dataset with deeply sequenced genes and many cells (top 72 genes of brain\_1.3m, Fig. 2a). See details of the subsampling procedure in Subsampling experiment in Supplementary Note 6). Third, a more controlled simulation that assumes the Poisson model was conducted to provide a more comprehensive evaluation (Supplementary Fig. 4). Both simulations exhibit similar qualitative behaviors and imply that the optimal sequencing depths  $n_{reads,g}^*$  for estimating different quantities are between 0.2 and 1. Therefore, we reached the conclusion that the optimal budget allocation for a single gene is to have  $\sim 1$  UMI per cell on average.

When there are many genes of primary biological interest, the gene among them with the smallest relative mean expression level becomes the bottleneck, as it has the fewest number of reads on average (Fig. 1c, top). We call



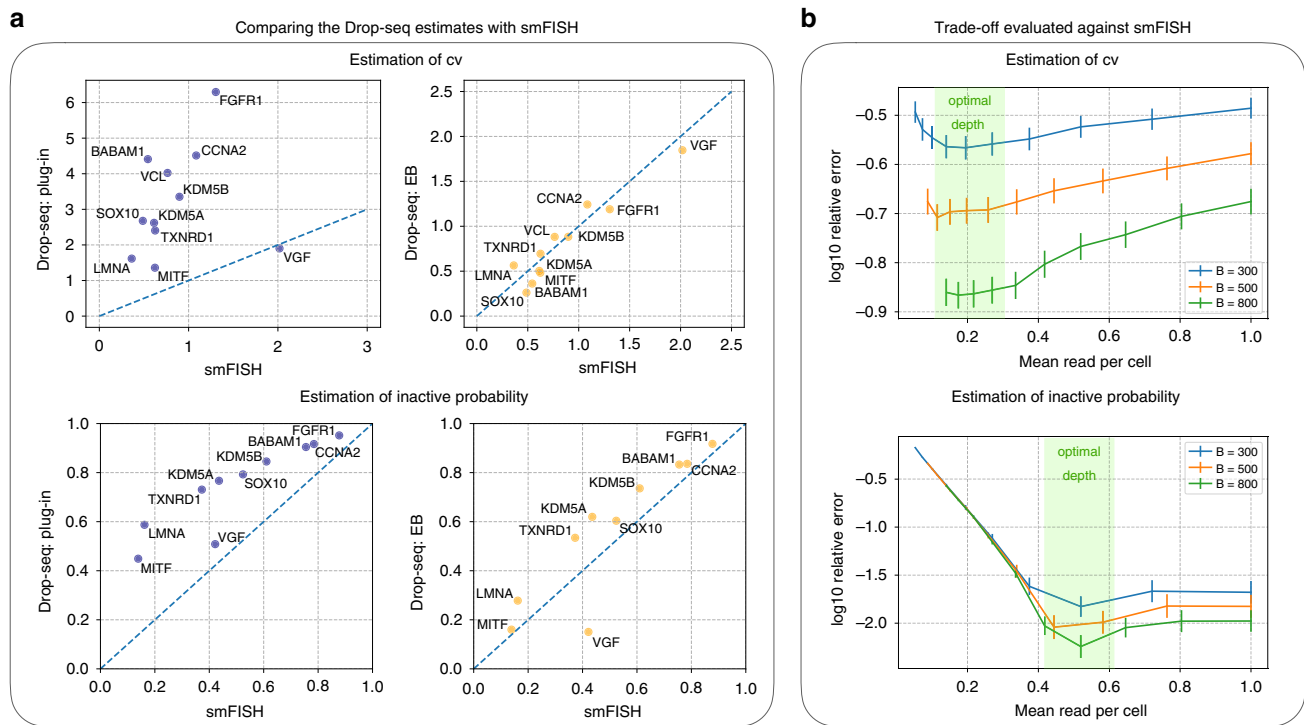
**Fig. 4 Gene module and gene network analysis.** **a** Top: the EB-estimated Pearson correlation for some marker genes in pbmc\_4k are visualized, ordered by different cell populations (top). The clear block-diagonal structure implies that the EB estimator is capable of capturing the gene functional groups. As a comparison, the plug-in estimator also recovers those modules but with a weaker contrast (bottom left panel, plug-in with 100%). Bottom: a subsample experiment further shows that the EB estimator can recover the module with 5% of the data. For the plug-in estimator, the first block (T cells) is blurred with 25% of the data, and the entire structure vanishes with 10% of the data. **b** Gene network based on the EB-estimated Pearson correlation using the pbmc\_4k dataset. Most gene modules correspond to important cell types or functions, including T cells, B cells, NK-cells, myeloid-derived cells, megakaryocytes/platelets, ribosomal protein genes, and mitochondrially encoded protein-coding genes. **c** Left: the estimated Pearson correlations between all genes and *LCK* (1st panel) and *CD3D* (2nd panel), two known T-cell markers. There are three modes for the EB-estimated values, where the positive mode, the zero mode, and the negative mode correspond to genes in the same module, different modules, and irrelevant genes, respectively. The plug-in estimated values are nonetheless much closer to zero even for the truly correlated ones, indicating an artificial shrinkage of the estimated values. Right: two instances where the EB estimates are significantly different from the plug-in estimates. The axes represent read counts, and the color codes the number of cells. Both gene pairs are biologically validated (see Gene network analysis in Methods). See also Supplementary Figs. 11–12 for more examples.

its relative mean expression level  $p^*$  the reliable detection limit, below which the estimation performance cannot be guaranteed. The optimal sequencing depth for the entire experiment  $n_{reads}^*$  is chosen so that the gene at the reliable detection limit has one read per cell (in expectation), which minimizes the worst-case error for all genes of interest. Compared with this optimal allocation, deeper sequencing (green) gives a homogeneous error across genes but at a much higher level, whereas a shallower sequencing (blue) gives a small error for a few highly expressed genes but its performance quickly deteriorates (Fig. 1c, bottom).

The recommended budget allocation in general suggests a slightly deeper sequencing depth as compared with existing datasets, e.g., 7k UMIs per cell for the pbmc\_4k dataset considering *MS4A1* and 14k UMIs per cell for the brain\_9k dataset considering *S100a10* (Supplementary Fig. 2). Such a depth is feasible for the current 10x Genomics' technology, which is estimated to be able to sequence 10–45k UMIs per cell where the actual values depend on different tissues

(Feasibility of the recommended sequencing depth in Supplementary Note 3). In addition, under such a sequencing depth, all analyses are valid as the Poisson model is still a good approximation of the sequencing process. Regarding the rare genes, since the UMI efficiency for the 10x technology is estimated to be 10–15%, in order to achieve one read per cell, the gene needs to have at least  $1/0.15 \approx 7$  transcripts in the cell. The gene *CD4* (Supplementary Fig. 2b) seems to be below this limit. For such genes, the recommendation should be sequencing until saturation.

The input parameter to the proposed experimental design approach, i.e., the detection limit  $p^*$ , corresponds to the smallest mean expression level among the list of genes of interest. Therefore, to carry out the proposed experimental design procedure, it is important to have an estimate of the mean expression levels for these genes. Such information may come from various sources whose data closely matched the system under study. First, researchers usually conduct pilot experiments before conducting the main experiment; the data from the pilot



**Fig. 5 Validation using smFISH data.** **a** The estimated CV (top) and inactive probability (bottom,  $\kappa = 2.5n_{\text{reads}}$ ) from the Drop-seq data are compared with the smFISH results. The EB estimates (right) are consistent with the smFISH results while there is a clear inflation for the plug-in estimates. **b** The sequencing budget tradeoff for estimating CV (top) and inactive probability (bottom,  $\kappa = 1$ , i.e., estimating the zero proportion at 1 read per cell) for the gene *MITF*. The relative error is evaluated against the gold standard smFISH result. 3-std confidence intervals are provided.

experiment can be used to provide such an estimate. Also, data from past studies or public databases, either scRNA-seq or bulk RNA-Seq, can be used to provide the estimate. Some popular databases include Tabula Muris (scRNA-seq)<sup>39</sup> for different mouse tissues, Human Cell Atlas<sup>29</sup> (scRNA-seq) and GTEx<sup>40</sup> (bulk RNA-Seq) for different human tissues, TCGA<sup>41</sup> (bulk RNA-Seq) for human cancer data, and GEO<sup>42</sup> (bulk/single-cell RNA-seq) for past studies. One caveat here is that different datasets may have different covariate compositions, like sex, age, or demographic factors. To evaluate the sensitivity of using reference data to estimate  $p^*$  for the proposed experimental design procedure, we consider four different types of reference data in Supplementary Figs. 18–19: in-sample bulk RNA-Seq or scRNA-seq, where the corresponding reference data were obtained from the same biological sample as the data for the current study, and their out-of-sample counterparts obtained from independent biological replicates. Our results suggest that although all four types of reference data can be used to determine the optimal sequencing depth accurately, in-sample scRNA-seq and out-of-sample bulk RNA-Seq should be considered as the most and least preferable sources of reference data respectively.

In practice, there is enough experimental flexibility to choose both the total sequencing budget  $B$  as well as the total number of cells  $n_{\text{cells}}$  to achieve the recommended allocation. The budget  $B$  is typically specified in terms of the total number of lanes that will be used for sequencing and is directly proportional to the sequencing cost of the experiment. For example, the 10x Genomics’ pbmc\_4k dataset was sequenced on one Illumina HiSeq4000 lane yielding a total of ~350 million reads, whereas the brain\_1.3m data set was sequenced on 88 HiSeq4000 lanes (11 flow cells) yielding ~30 billion reads. Sample multiplexing can also be utilized to achieve fractional lane occupancies for smaller experiments. Now, given a fixed budget  $B$ , one can adjust the desired sequencing depth ( $n_{\text{reads}} = B/n_{\text{cells}}$ ) by selecting the total number of cells at the library preparation stage of the experiment. Although all single-cell RNA-seq assays rely on the Illumina platform for sequencing, the library preparation stage (e.g., single-cell isolation, mRNA capture, and barcoding) is technology-specific<sup>22–24,28,43</sup>. Nevertheless, it is possible to accurately choose the total number of cells by adjusting the cell concentration (cells/ $\mu\text{l}$ ) and the final cell suspension volume that is going to be used in the process. For example, the 10x Genomics Chromium platform can be adjusted to yield from 500 to 10K cells per lane in a single run (10x user manual: <https://support.10xgenomics.com/permalink/3vzDu3zQjY0o2AqkkkI4CC>). For larger experiments, multiple lanes can be used (e.g., the brain\_1.3m dataset was prepared on 133 10x Genomics Chromium lanes, each optimized to capture ~10k cells). Even though the library preparation stage can incur additional costs for a single-cell RNA-seq experiment, these costs are independent of the sequencing process, can vary significantly across different technologies<sup>44</sup>, and are in general decreasing rapidly.

**Empirical Bayes estimators.** The EB estimators refer to the estimators that are aware of the noise model (which is Poisson here) and correct for the noise introduced by it. As they estimate the prior gene distribution  $P_X$  in the hierarchical model (2) from the observed data  $Y_c$ , sometimes by estimating the moments of the prior distribution  $P_X$ , they are usually associated with the names empirical Bayes, moment matching, or density deconvolution. Here, we use the term EB to refer to them in general.

As an illustrating example, consider a simplified model that for cell  $c$  and gene  $g$ :

$$X_{cg} \sim P_{X_g}, Y_{cg}|X_{cg} \sim \text{Poi}(X_{cg}).$$

The plug-in estimator estimates the gene variance by the sample variance of UMIs, i.e.,

$$\widehat{\text{var}}_g^{\text{plug-in}} = \frac{1}{n_{\text{cells}} - 1} \sum_{c=1}^{n_{\text{cells}}} (Y_{cg} - \bar{Y}_{cg})^2,$$

where  $\bar{Y}_{cg} = \frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} Y_{cg}$  is the empirical mean. However, the estimated value is usually overly variable owing to the presence of the Poisson noise. Indeed,

$$\mathbb{E}[\widehat{\text{var}}_g^{\text{plug-in}}] = \text{Var}[Y_{cg}] = \text{Var}[X_{cg}] + \mathbb{E}[X_{cg}],$$

where the second term  $\mathbb{E}[X_{cg}]$  corresponds to the technical variation introduced by the Poisson noise. Then, conceptually we can write:

$$\text{plug-in variance} = \text{biological truth} + \text{Poisson noise},$$

from which we can see that the plug-in estimate is inflated by the Poisson noise. In this case, this bias can be easily corrected by simply subtracting the mean, and the corresponding EB variance estimator can be written as

$$\widehat{\text{var}}_g^{\text{EB}} = \frac{1}{n_{\text{cells}} - 1} \sum_{c=1}^{n_{\text{cells}}} (Y_{cg} - \bar{Y}_{cg})^2 - \frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} Y_{cg}.$$

The EB estimators considered in the paper are listed in Table 1, along with the plug-in estimators for comparison. In literature, they are designed in a case-by-case fashion<sup>16–18,24,45–47,47–50</sup> (more details in Supplementary Note 4).

**Empirical evaluation of the tradeoff.** We conducted two sets of simulations to evaluate the estimation error under different budget splits, which differ in how the data are generated. The first simulation (Fig. 2a) subsampled from a high-budget dataset consisting of the top 72 genes from the brain\_1.3m dataset. These genes were chosen because they have at least 10 reads per cell, providing a deep dataset to



**Table 1 Comparison of the plug-in estimator and the EB estimator.**

	plug-in	EB
1st moment $M_{1,g}$	$\frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} \frac{Y_{cg}}{y_c n_{\text{reads}}}$	same
2nd moment $M_{2,g}$	$\frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} \frac{Y_{cg}^2}{(y_c n_{\text{reads}})^2}$	$\frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} \frac{Y_{cg}^2 - Y_{cg}}{(y_c n_{\text{reads}})^2}$
kth moment $M_{k,g}$	$\frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} \frac{Y_{cg}^k}{(y_c n_{\text{reads}})^k}$	$\frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} \frac{\prod_{i=0}^{k-1} (Y_{cg} - i)}{(y_c n_{\text{reads}})^k}$
1st pairwise moment $M_{11,g_1,g_2}$	$\frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} \frac{1}{n_{\text{reads}}^2} Y_{cg_1} Y_{cg_2}$	same
Inactively probability $p_{0,g}(\kappa)$	$\frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} \mathbb{I}\{Y_{cg}=0\}$	$\frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} a_{Y_{cg}}$
Pairwise inactive probability $p_{0,g_1,g_2}(\kappa)$	$\frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} \mathbb{I}\{Y_{cg_1}=Y_{cg_2}=0\}$	$\frac{1}{n_{\text{cells}}} \sum_{c=1}^{n_{\text{cells}}} a_{Y_{cg_1}} a_{Y_{cg_2}}$
Distribution $P_{X_g}$	Empirical distribution of $Y_{cg}$ (scaled by $1/n_{\text{reads}}$ )	$\hat{P}_{X_g}$ that most likely gives empirical distribution of $Y_{cg}$ via model (2)

The two estimators are written in similar forms for better comparison. For the inactive probability (and the pairwise case),  $a_{Y_{cg}}$  is a coefficient that depends on  $Y_{cg}$ ,  $\kappa$ , and  $n_{\text{reads}}$ . See Inactive probability in Supplementary Note 5 for the exact expression and other details.

perform the subsample experiments. This simulation better matches the real data as the subsampling procedure does not assume the Poisson model (see Sub-sampling experiment in Supplementary Note 6). However, as we did not know the true gene distribution, the plug-in estimates of the high-budget dataset that we subsample from were used as proxies of the ground truth, against which we evaluated the estimation error. The second simulation, corresponding to Supplementary Fig. 4, generated the data according to model (2), where the true gene distribution  $P_X$  was obtained by using the empirical distribution of the first 100 highly expressed genes in the pbmc\_4k dataset. This setting better validates the theory as it assumes the same model. Moreover, the estimation error is exact as the ground truth is available. Both simulations include many genes to address the heterogeneity of the gene distribution, and the genes considered here, being top genes in the dataset, have similar mean expression levels so that the mean reads over all genes can well represent the mean reads for each gene. Both simulations exhibit qualitatively the same behavior, validating the theory that the optimal depth (mean reads per cell per gene) is a constant that does not depend on the budget.

**Post hoc guidance for reliable estimation.** The feasible region (top) and the post hoc table (bottom) were obtained via simulation, where we fixed the number of cells (1k, 5k, 10k, 30k, 70k) and studied how the error decreases as a function of the sequencing depth (Supplementary Figs. 5–6). The data were generated according to model (2) similar to the second tradeoff simulation, where the empirical distributions of the marker genes in pbmc\_4k and brain\_9k were used as the true gene distribution, respectively, to account for heterogeneity in different tissues. The true gene distribution was normalized so that each gene has the same mean expression level. As a result, the mean reads over all genes were the same as mean reads for each gene, providing a single-gene level error characterization. The post hoc table was obtained by finding the smallest sequencing depth such that the relative error was smaller than 0.1 (−2 in the log10 scale for the relative squared error and −1 for other errors, see Definition of errors in simulations in Supplementary Note 6). The results for both simulations were qualitatively the same. Hence, only the table for pbmc\_4k was included.

**Comparing the performance of plug-in and EB estimators.** Figure 3a demonstrates that the EB estimator is adaptive to different sequencing depths while the plug-in estimator is not. The top panel shows the estimated CV using the plug-in and EB estimators under different sequencing depths, where we can see clear inflation for the plug-in estimates. The full data are from pbmc\_4k, and the subsample rate ranges from 0.2 to 1 (1 corresponds to the full data). The experiment was repeated five times, and the 3-std confidence interval was provided. The results for other genes, as well as for estimating the inactive probability, can be found in Supplementary Fig. 7. The middle panel compares the estimated CV from two datasets of the same tissue. Genes with at least 0.1 reads per cell were considered as our post hoc analysis showed that CVs of genes below this level could not be reliably estimated. The EB estimator may produce an invalid result when the plug-in variance is smaller than the plug-in mean of a gene, which was not accounted for by the Poisson model. Such cases were not common and were excluded while counting the number of genes that are above or below the red line. Hence, the total number of genes of the two panels may slightly differ. More results are in Supplementary Figs. 8–9. The bottom panel shows that the EB estimator can recover the gene distribution from shallow sequencing data. The shallow data were generated by subsampling to have 20% reads of the full data. For error evaluation, the recovered distribution was rescaled to have the same mean as the empirical distribution from the full data. See Supplementary Fig. 10 for more results.

Figure 3b investigates the common task where the most informative features (genes) were selected based on CV, and PCA was then performed on the selected features. The data were from pbmc\_4k and was clipped at the 99th quantile to remove outliers. Such a procedure was also used in previous works on applying PCA to scRNA-seq data<sup>11</sup>. The top 500 genes with the highest CV were selected and the

PCA scores were plotted for the 2nd and 3rd PC direction. The first direction was skipped because it corresponded to the variation in cell sizes. The results on the full data and the subsampled data (three times shallower) were compared, showing that the EB estimator is more consistent than the plug-in estimator.

Figure 4a considers recovering gene functional groups using Pearson correlation. We used the pbmc\_4k dataset here as the biological structure of the PBMCs is well-understood. The major cell populations identified in this dataset are T cells (*IL7R*, *CD3D/E*, *LCK*), NK-cells (*NKG7*, *PRF1*, *KLRD1*, *GZMA*, *HOPX*, *CST7*), B cells (*CD79A*, *BANK1*, *IGHD*, *LINC00926*, *MSA11*), myeloid-derived cells (*S100A8/9*, *MNDA*, *FGL2*, *CLECTA*, *IFI30*) and megakaryocytes/platelets (*PF4*, *PPBP*). The heatmap of the EB-estimated Pearson correlation of those genes were visualized in Fig. 4a top, which shows that the EB estimator can well capture the gene functional groups. A subsample experiment was then conducted to investigate how well the estimators can recover the modules from the shallow sequencing data. The data were subsampled from the full data with rates 100% (full data), 25%, 10%, and 5%. The EB estimator can recover the module at a much shallower depth as compared to the plug-in estimator.

**Gene network analysis of the pbmc\_4k dataset.** The gene network (Fig. 4b) was constructed based on the EB-estimated Pearson correlation using the pbmc\_4k dataset. The genes were filtered to have the EB-estimated variance larger than 0.1, resulting in 791 genes in total. A correlation larger than 0.8 was considered as a gene–gene edge. We found that varying the threshold from 0.4 to 0.95 did not significantly alter the result. The gene modules were identified based on knowledge of marker genes and gene pathways, as well as previous studies on PBMCs (see Gene module identification in Supplementary Note 6). We also note that the existence of megakaryocytes/platelets may be due to the imperfection of PBMC isolation, and since many genes were expressed in multiple cell populations (e.g., *CD74*, *CD27*), the resulting annotation only gives a rough picture of the underlying gene functional groups.

Next, we considered some important genes and plot their correlations with all other genes (Fig. 4c left, Supplementary Fig. 11). As a general phenomenon, the EB-estimated values are more spread out and exhibit different modes corresponding to genes that interact differently with the gene of interest. The plug-in estimated values are nonetheless much closer to zero even for genes that are known to be well-correlated.

Finally, we considered the gene pairs where the estimated values for the EB estimator and the plug-in estimator differ significantly (>0.7). Out of 1054 such pairs, 91 were also annotated based on STRING<sup>51</sup>, yielding a *p* value of 4.2e-11 while testing against the null hypothesis that the gene pairs were selected at random based on a one-sided hypergeometric distribution test (see Gene module identification in Supplementary Note 6). We plot the histograms of several such pairs and show that all of them have clear biological interpretations (Fig. 4c right, Supplementary Fig. 12). *LY86* (also known as *MD1*) is a secreted protein that has been shown to have an important role in T-cell activation, whereas *CD3E* is expressed within T cells (see Gene module identification in Supplementary Note 6). These two genes are not co-expressed and hence, are negatively correlated. *POMP* encodes a chaperone for proteasome assembly, whereas *PSMA7* is one of the 17 essential subunits for the complete assembly of the 20S proteasome complex. Hence, the two genes work together for proteasome assembly and should be positively correlated. The EB-estimated correlation is one and is probably an over-estimate owing to the randomness of the estimator. However, the actual Pearson correlation should not be much smaller than 1. In spite of the strong biological evidence, the plug-in estimator gives very small values owing to the presence of sequencing noise (See also Supplementary Fig. 12).

**smFISH experiments for validation.** For validation, we considered two datasets, where both the scRNA-seq and the smFISH data are available. smFISH can be

regarded as the gold standard for measuring the number of mRNAs in a cell and was used as a proxy for the ground truth (see Details of the smFISH experiments in Supplementary Note 6 for more details).

In the first dataset, both Drop-seq and smFISH were applied to the same melanoma cell line<sup>15</sup>. A total of 5763 cells and 12,241 genes were kept for analysis from the Drop-seq experiment, with a median of 1473 UMIs per cell. Of these genes, 24 were also profiled using smFISH. We further excluded genes with zero-UMI count in >97% of the cells and one more gene, *FOSL1*, owing to its abnormal behavior (*FOSL1* was also excluded in a recent work<sup>16</sup> analyzing the dataset). We considered two distributional quantities, CV and inactive probability (with  $\kappa = 2.5n_{\text{reads}}$ ), where we note that the latter has the interpretation of the proportion of zeros when the data were sequenced 2.5 times deeper. We compared the plug-in and the EB-estimated results from the Drop-seq data against the corresponding result from the smFISH data in Fig. 5a, where the smFISH estimates can be considered as the ground truth. Here, the gene *VCL* was omitted in the experiment of estimating the inactive probability because the corresponding smFISH data do not have enough cells to subsample from (4691, fewer than the number of cells captured by Drop-seq, which is 5763). The consistency between the EB estimates and the smFISH result indicates that the EB estimates are close to the ground truth. Furthermore, we investigated the optimal sequencing depth (Figure 5b and Supplementary Fig. 14) by fixing the budget and varying the sequencing depth, where the error was evaluated against the gold standard smFISH result. As this was done by subsampling from the original dataset, to ensure a wide range, only two genes with relatively more reads (*MITF* and *VGF*) were considered. Figure 5b and Supplementary Fig. 14 are qualitatively similar to the simulation results in Fig. 2a and Supplementary Fig. 4, showing an optimal depth between 0.1 and 0.6. This is consistent with our previous experiments based on 10x Genomics' data and the experimental design guidelines we provide in this work, i.e., that the optimal depth for estimating different quantities is 0.2–1 read per cell per gene.

In the second dataset, both CEL-seq and smFISH were applied to the same mESC cell line and culture conditions<sup>27</sup> (smFISH data from D. Grün, personal communication). Again, the plug-in and the EB-estimated results from the CEL-seq data were compared against the corresponding result from the smFISH data in Supplementary Fig. 13 for nine genes measured by smFISH, where we observed a good consistency between the EB and the smFISH results. As there are only 80 cells, we did not perform the subsampling experiment for this dataset.

Overall, the comparisons between the scRNA-seq and the smFISH results imply that our model matches the real data well, and the proposed EB estimator is able to provide estimates that are close to the ground truth. Also, the subsampling experiments in Fig. 5b and Supplementary Fig. 14 indicate that the optimal depth, evaluated using the smFISH data, is consistent with the main claim of the paper.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The 10x datasets were generated by 10x Genomics' v2 chemistry<sup>22</sup>. They are publicly available and can be downloaded via the following links:

pbmc\_4k: <https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k>

pbmc\_8k: <https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc8k>

brain\_1k: [https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/neurons\\_900](https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/neurons_900)

brain\_2k: [https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/neurons\\_2000](https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/neurons_2000)

brain\_9k: [https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/neuron\\_9k](https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/neuron_9k)

brain\_1.3m: [https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M\\_neurons](https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons)

293T\_1k, 3T3\_1k: [https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm\\_1k](https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm_1k)

293T\_6k, 3T3\_6k: [https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm\\_6k](https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm_6k)

293T\_12k, 3T3\_12k: [https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm\\_12k](https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm_12k)

We note that pbmc\_4k and pbmc\_8k are from the same donor; brain\_1k and brain\_9k are also from the same donor. Also, the following pairs of datasets are sequenced together: 293T\_1k and 3T3\_1k, 293T\_6k and 3T3\_6k, 293T\_12k and 3T3\_12k. These six datasets are from the same biological sample.

The Drop-seq dataset and the corresponding smFISH data can be found from the original paper<sup>15</sup> or a recent paper that analyzed the dataset<sup>16</sup>. The CEL-seq data can be found from the original paper<sup>27</sup>. The smFISH data accompany the CEL-seq can be obtained by contacting the author. The three ERCC datasets (Zheng, Klein, Svensson) can be found in a recent paper that analyzed the data set<sup>16</sup>, where we have used the  $2 \times$  (control RNA + ERCC) data in the Svensson et al.<sup>52</sup> paper. The Klein dataset with the pure RNA controls (the Klein ERCC dataset being part of it) can be found from the original paper<sup>24</sup>. The data for sensitivity analysis (Supplementary Figs. 18–19) can be found from the original paper<sup>53</sup>.

## Code availability

We developed the python package sceb (single-cell empirical Bayes) for the EB estimators used in this paper (available on PyPI). The code to reproduce all experiments and generate the figures presented in this paper can be found at [https://github.com/martinjzhang/single\\_cellEb](https://github.com/martinjzhang/single_cellEb).

Received: 6 September 2018; Accepted: 13 December 2019;

Published online: 07 February 2020

## References

- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
- Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).
- Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599 (2018).
- Streets, A. M. & Huang, Y. How deep is enough in single-cell RNA-seq? *Nat. Biotechnol.* **32**, 1005 (2014).
- Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17**, 63 (2016).
- Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).
- Dal Molin, A. & Di Camillo, B., How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. *Brief. Bioinform.* **20**, 1384–1394 (2018).
- Ecker, J. R. et al. The brain initiative cell census consortium: lessons learned toward generating a comprehensive brain cell atlas. *Neuron* **96**, 542–557 (2017).
- Pollen, A. A. et al. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053 (2014).
- Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
- Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Systems* **2**, 239–250 (2016).
- Shalek, A. K. et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363 (2014).
- Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
- Rizzetto, S. et al. Impact of sequencing depth and read length on single cell RNA sequencing data of t cells. *Sci. Rep.* **7**, 12781 (2017).
- Torre, E. et al. Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA fish. *Cell Syst.* **6**, 171–179 (2018).
- Wang, J. et al. Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl Acad. Sci.* **115**, E6437–E6446 (2018).
- Efron, B. Two modeling strategies for empirical Bayes estimation. *Stat. Sci.* **29**, 285 (2014).
- Efron, B. Empirical Bayes deconvolution estimates. *Biometrika* **103**, 1–20 (2016).
- Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
- Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881–2887 (2007).
- Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G. & Chen, X. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol.* **19**, 70 (2018).
- Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
- Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
- Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
- Efron, B., *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, 1. Cambridge University Press, 2012.
- Huang, M. et al. Saver: gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15**, 539 (2018).
- Grün, D., Kester, L. & Van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637 (2014).
- Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. Cel-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).

29. Regev, A. et al. Science forum: the human cell atlas. *Elife* **6**, e27041 (2017).
30. Kim, J. K., Kolodziejczyk, A. A., Ilicic, T., Teichmann, S. A. & Marioni, J. C. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **6**, 8687 (2015).
31. Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136–138 (2009).
32. Korthauer, K. D. et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* **17**, 222 (2016).
33. Jolliffe, I. T., Principal component analysis and factor analysis, in *Principal component analysis*, 115–128, Springer, 1986.
34. Abid, A., Zhang, M. J., Bagaria, V. K. & Zou, J. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nat. Commun.* **9**, 2134 (2018).
35. Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 888–905 (2000).
36. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).
37. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
38. Mohammadi, S., Davila-Velderrain, J., Kellis, M. & Grama, A. DECODE-ing sparsity patterns in single-cell RNA-seq, Preprint at <https://doi.org/10.1101/241646v2> (2018).
39. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris: the tabula muris consortium. *Nature* **562**, 367 (2018).
40. Consortium, G. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204 (2017).
41. Weinstein, J. N. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113 (2013).
42. Edgar, R., Domrachev, M. & Lash, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
43. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
44. AlJanahi, A. A., Danielsen, M. & Dunbar, C. E. An introduction to the analysis of single-cell rna-sequencing data. *Mol. Ther. Methods Clin. Dev.* **10**, 189–196 (2018).
45. Jiao, J., Venkat, K., Han, Y. & Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Transact. Inform. Theory* **61**, 2835–2885 (2015).
46. Yang, Y. Wu et al. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Stat.* **47**, 857–883 (2019).
47. Orłitsky, A., Suresh, A. T. & Wu, Y. Optimal prediction of the number of unseen species. *Proc. Natl Acad Sci.* **113**, 13283–13288 (2016).
48. Kong, W. et al. Spectrum estimation from samples. *Ann. Stat.* **45**, 2218–2247 (2017).
49. Good, I. & Toulmin, G. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45–63 (1956).
50. Efron, B. & Thisted, R. Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* **63**, 435–447 (1976).
51. Szklarczyk, D. et al. String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2014).
52. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381 (2017).
53. Ding, J. et al., Systematic comparative analysis of single cell rna-sequencing methods, Preprint at <https://doi.org/10.1101/632216v2> (2019).

## Acknowledgements

This research was in part motivated by discussions on the experimental design question in the Human Cell Atlas First Annual Jamboree meeting. We thank Lior Pachter for his valuable input and constructive suggestions throughout the course of this study; Jase Gehring, Wenying Pan, and Taibo Li for their helpful feedback; and Dominic Grün for providing the smFISH data corresponding to the CEL-seq data. Thanks also to Patrick Marks for very useful feedback on an earlier version of the paper. D.T. and M.J.Z. are supported in part by the Center of Science of Information, an NSF Science and Technology Center, under grant agreement CCF-0939370 and in part by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG008164. M.J.Z. is also supported by a Stanford Graduate Fellowship (Inventec Fellow). V.N. is supported in part by the Center for Science of Information and in part by a gift from Qualcomm Inc.

## Author contributions

M.J.Z. and V.N. conceived the idea and performed the empirical experiments. M.J.Z. performed the theoretical analysis. M.J.Z., V.N. and D.T. wrote the manuscript. D.T. supervised the research. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-14482-y>.

**Correspondence** and requests for materials should be addressed to D.T.

**Peer review information** *Nature Communications* thanks Jay West and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020