Gene expression distribution deconvolution in single-cell RNA sequencing

Jingshu Wang[a], Mo Huang[a], Eduardo Torre[b], Hannah Dueck[c], Sydney Shaffer[b], John Murray[c], Arjun Raj[b], Mingyao Li[d], and Nancy R. Zhang[a,1]

[a]Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104; [b]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA 19104; [c]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104; and [d]Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA 19104

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved May 29, 2018 (received for review December 6, 2017)

Single-cell RNA sequencing (scRNA-seq) enables the quantification of each gene's expression distribution across cells, thus allowing the assessment of the dispersion, nonzero fraction, and other aspects of its distribution beyond the mean. These statistical characterizations of the gene expression distribution are critical for understanding expression variation and for selecting marker genes for population heterogeneity. However, scRNA-seq data are noisy, with each cell typically sequenced at low coverage, thus making it difficult to infer properties of the gene expression distribution from raw counts. Based on a reexamination of nine public datasets, we propose a simple technical noise model for scRNA-seq data with unique molecular identifiers (UMI). We develop deconvolution of single-cell expression distribution (DESCEND), a method that deconvolves the true cross-cell gene expression distribution from observed scRNA-seq counts, leading to improved estimates of properties of the distribution such as dispersion and nonzero fraction. DESCEND can adjust for cell-level covariates such as cell size, cell cycle, and batch effects. DESCEND's noise model and estimation accuracy are further evaluated through comparisons to RNA FISH data, through data splitting and simulations and through its effectiveness in removing known batch effects. We demonstrate how DESCEND can clarify and improve downstream analyses such as finding differentially expressed genes, identifying cell types, and selecting differentiation markers.

single-cell transcriptomics | RNA sequencing | differential expression | Gini coefficient | highly variable genes

Cells are the basic biological units of multicellular organisms. Within a cell population, individual cells vary in their gene expression levels, reflecting the dynamism of transcription across cells (1–5). Traditional microarray and bulk RNA-sequencing (RNA-seq) technologies profile the average gene expression level of all cells in the population. In contrast, recent single-cell RNA-seq (scRNA-seq) methods enable the quantification of a much richer set of properties of the gene expression distribution across cells. For example, measures of dispersion such as the coefficient of variation (CV) and the Gini coefficient can be used to elucidate biological states that are not reflected in the population average (6–8). Two-state mixture models, alternatively, allow a better understanding of transcriptional regulation at the single-cell level (5, 9–11).

However, it is challenging to compute such distribution-based statistics of true gene expression due to the technical noise in scRNA-seq data (12–16). Single-cell RNA-seq protocols are complex, involving multiple steps each contributing to the substantially increased noise level of scRNA-seq relative to bulk RNA-seq. Unique molecular identifiers (UMI) (17) were introduced as a barcoding technique to reduce amplification noise, but the observed expression distribution computed from observed UMI counts is, for most genes, still a poor representation of their true expression distribution.

Recently, many computational methods for scRNA-seq analysis have been proposed, including methods for quantifying dispersion, for characterizing expression "states" using mixture models, and for finding differentially expressed genes (6, 18–26). Although some of these methods have taken technical noise into consideration, to our knowledge, there is currently no method for recovering the cross-cell gene expression distribution from scRNA-seq data, unless strong assumptions are made about this distribution. There is also a lack of methods for comparing aspects of this true biological distribution beyond the mean, especially when there is a need to adjust for confounding factors. In fact, there is still active debate on the technical noise distribution for scRNA-seq data, and a proper technical noise model is critical for studying the underlying distribution.

Here we develop deconvolution of single-cell expression distribution (DESCEND), a statistical method that deconvolves the true cross-cell gene expression distribution from observed scRNA-seq counts and quantifies how this distribution depends on cell-level covariates. Examples of common cell-level covariates are cell size, defined as the total number of RNA molecules in a cell, and cell-cycle stage. DESCEND adopts the

Significance

We developed deconvolution of single-cell expression distribution (DESCEND), a method to recover cross-cell distribution of the true gene expression level from observed counts in single-cell RNA sequencing, allowing adjustment of known confounding cell-level factors. With the recovered distribution, DESCEND provides reliable estimates of distribution-based measurements, such as the dispersion of true gene expression and the probability that true gene expression is positive. This is important, as with better estimates of these measurements, DESCEND clarifies and improves many downstream analyses including finding differentially expressed genes, identifying cell types, and selecting differentiation markers. Another contribution is that we verified using nine public datasets a simple "Poisson-alpha" noise model for the technical noise of unique molecular identifier-based single-cell RNA-sequencing data, clarifying the current intense debate on this issue.

Author contributions: J.W. and N.R.Z. conceptualized the study; J.W. formulated the model, developed the algorithms and methods, and executed the simulations and real data analysis; J.W., M.L., and N.R.Z. planned the model validation and case studies; E.T., H.D., S.S., J.M., and A.R. provided the RNA FISH and Drop-Seq data of the same melanoma cell line; and J.W., M.H., M.L., and N.R.Z. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: nzh@wharton.upenn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1721085115/-/DCSupplemental.

Published online June 26, 2018.

"G-modeling" distribution deconvolution framework (27), which avoids constraining parametric assumptions. The accuracy of DESCEND is evaluated using population-matched RNA fluorescence in situ hybridization (FISH) data, in silico sample splitting, and parametric and down-sampling simulations. Our evaluations show that, under very reasonable data quality assumptions, DESCEND can accurately deconvolve the true gene expression distribution, leading to improved estimates of expression dispersion and proportion of cells with zero/nonzero expression. We benchmark against existing methods (6, 25, 28) for analyses in which comparable methods exist, but focus on demonstrating the unique applications of DESCEND in the case studies that follow.

Although the DESCEND framework can be used with any technical noise model, the datasets we use in this paper all use UMIs, for which we have decided upon a satisfactory noise model. There is a lot of debate regarding what noise model to use, even for UMI-based scRNA-seq data. Through a reanalysis of nine public datasets, we show that a Poisson distribution is sufficient to capture the technical noise in single-cell UMI counts, once the cross-cell differences in capture and sequencing efficiency have been accounted for. Thus, DESCEND adopts a "Poisson-alpha" noise model for single-cell UMI counts, which allows fast computation and stable estimation. We show using the data from Tung et al. (29) that, with this noise model, DESCEND can effectively remove artificial differences between known experimental batches.

We demonstrate the applications of DESCEND through three case studies. The first case study illustrates how to conduct differential expression analysis under a two-state model for gene expression. Here, we also conduct a transcriptome-wide examination of how gene expression distributions are associated with cell size, again using population-matched RNA FISH to validate our findings. Next, we illustrate how to use the DESCEND estimated dispersion measures, and their standard errors, for downstream inference. The second case study focuses on cell type identification, and the third case study focuses on the selection of marker genes for cell differentiation.

Results

Model Overview. Fig. 1 gives an overview of the DESCEND framework. The observed counts in an scRNA-seq experiment are a noisy reflection of true expression levels. We model the observed count Y_{cg} for gene g in cell c as a convolution of the true gene expression λ_{cg} and technical noise,

$$Y_{cg} \sim F_{cg}(\lambda_{cg}), \quad \lambda_{cg} \sim H_g(\cdot),$$

where $F_{cg}(\cdot)$ quantifies all aspects of the technical noise introduced in the experiment and H_g represents the true underlying expression distribution of gene g across cells. DESCEND deconvolves H_g from the noisy observed counts Y_{cg}, thus allowing for the estimation of any quantity related to the true underlying biological distribution. One difference between DESCEND and existing methods is that DESCEND models H_g using a spline-based exponential family, which avoids restrictive parametric assumptions while allowing the flexible modeling of dependence on cell-level covariates.

Currently, the noise model used by DESCEND is designed for single-cell experiments that use UMIs. For extension to non-UMI read counts, see *Discussion*. For UMI-based single-cell RNA-seq data, DESCEND uses the default noise model $Y_{cg} \sim F_c(\lambda_{cg}) = \text{Poisson}(\alpha_c \lambda_{cg})$, where α_c is a cell-specific scaling constant. This model was suggested by ref. 14, and in the next section, we show through a reexamination of public data that this model is sufficient for capturing the technical noise in UMI counts when there are no batch effects. To account for batch effects, DESCEND allows a more complicated model, $Y_{cg} \sim F_c(\lambda_{cg}) = \text{Poisson}(\alpha_{cg} \lambda_{cg})$.

In the default noise model, DESCEND sets α_c, by default, to the total UMI barcode count of cell c,

$$M_c = \sum_g Y_{cg}, \qquad [1]$$

where the summation is over all biological genes (i.e., excluding spike-ins). This leads to the interpretation of λ_{cg} being the

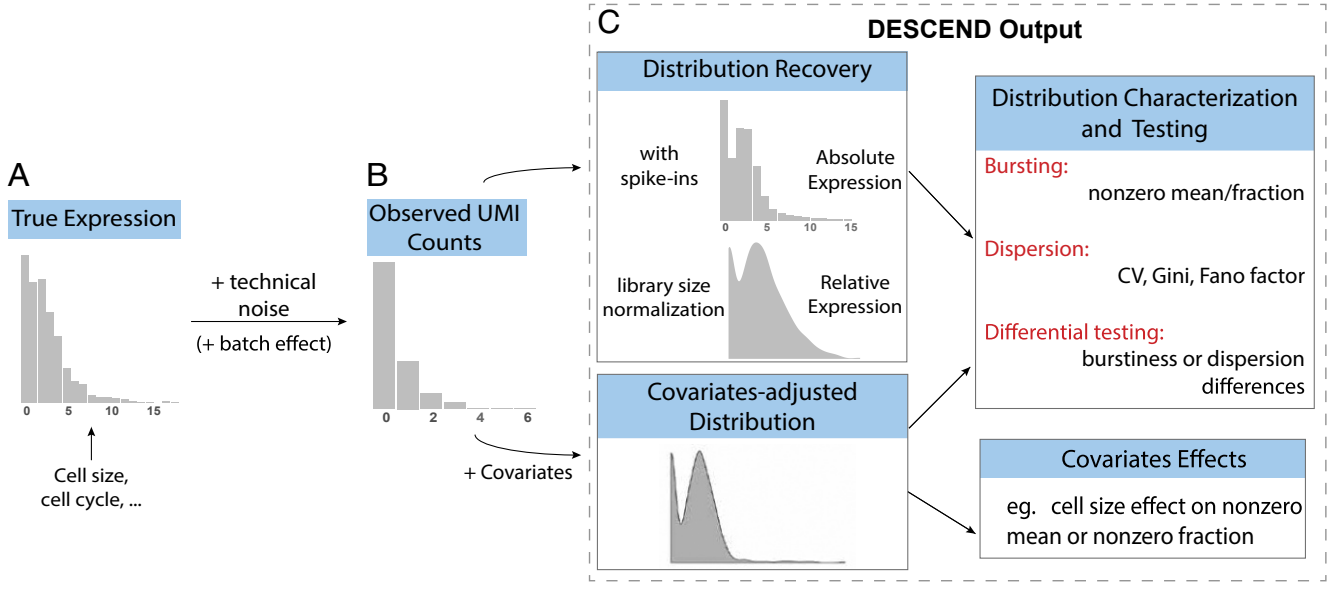


Fig. 1. Illustration of the framework. (*A* and *B*) The cross-cell distribution of observed counts Y_{cg} (*B*) is assumed to be a convolution of the distribution of true gene expression (*A*) and technical noise. (*C*) For each gene, the output of DESCEND includes the distribution of the absolute expression levels when spike-ins are available, the distribution of relative expression with library size normalization, the distribution of covariates-adjusted expression level if covariates are presented, estimates of the bursting and dispersion parameters, differential testing results comparing the change between two cell populations, and the effects of observed covariates on gene expression.

relative expression of gene g in cell c. If, instead, we were interested in deconvolving the distribution of the absolute gene expression count, then we would need to rely on cell-specific spike-ins to compute the efficiency, defined as the proportion of transcripts in the cell that are sequenced,

$$\text{efficiency of cell } c = \sum_{g \text{ is spike-in}} Y_{cg} / \sum_{g \text{ is spike-in}} \lambda_g, \qquad [2]$$

where λ_g is the expected input molecule count of spike-in gene g, computable from known dilution ratios. Setting α_c to this estimated efficiency of cell c leads to the interpretation of λ_{cg} being the absolute expression of gene g in the cell. Details are in *Materials and Methods* and *SI Appendix*, *Mathematical Details of DESCEND*.

The true gene expression distribution H_g is expected to be complex, owing to the possibility of multiple cell subpopulations and to the transcriptional heterogeneity within each subpopulation. In particular, this distribution may have several modes and an excessive amount of zeros and cannot be assumed to abide by known parametric forms. To allow for such complexity, DESCEND adopts the technique from Efron (27) and models the gene expression distribution as a zero-inflated exponential family which has the zero-inflated Poisson, lognormal, and Gamma distributions as special cases. Natural cubic splines are used to approximate the shape of the gene expression distribution (*Materials and Methods*).

One meaningful aspect of H_g is the proportion of cells where the true expression of the gene is nonzero; that is,

$$\text{nonzero fraction} \triangleq \mathbb{P}[\lambda_{cg} \neq 0]. \qquad [3]$$

Complementary to this is the nonzero mean, defined as the average expression level among cells where the gene is expressed,

$$\text{nonzero mean} \triangleq \mathbb{E}[\lambda_{cg} \mid \lambda_{cg} \neq 0]. \qquad [4]$$

Note that Eqs. **3** and **4** refer to the underlying, unobserved, true gene expression distribution. The concepts of nonzero fraction and nonzero mean have appeared, under varying definitions and differing names, in single-cell studies (5, 25, 30), yet many existing approaches to estimate them (5, 18, 30–32) do not account for technical noise. If the population from which the cells are sampled can be assumed to be ergodic, then a two-state transcriptional bursting model (9, 11, 33), formulated as a periodic stochastic dynamic process, leads to a Poisson-Beta distribution for gene expression across cells. In that scenario, Eqs. **3** and **4** can be derived from the burst frequency and burst size parameters defined in the Poisson-Beta distribution. However, the strong ergodicity assumption is, in most cases, too ideal for scRNA-seq experiments, in which cell populations are unavoidably heterogeneous even when limited to a specific cell type. In DESCEND, we choose not to assume the Poisson-Beta distribution and instead focus on the more transparent quantities [**3**, **4**].

DESCEND allows the modeling of covariate effects on both the nonzero fraction and nonzero mean. When covariates are specified, DESCEND uses a log-linear model for the covariate effect on nonzero mean and a logit model for the covariate effect on nonzero fraction. In this case, $\lambda_{cg} \sim H_{cg}$; that is, the distribution of λ_{cg} is cell specific, and the deconvolution result is the covariate-adjusted expression distribution (*Materials and Methods*).

A cell-level covariate that we study in detail is the cell size, defined as the total RNA molecule count in the cell. Cell size can be estimated by spike-in molecules added to the cell after RNA extraction: Let α_c be the efficiency of cell c obtained through Eq. **2**; then

$$\text{size estimate of cell } c = M_c/\alpha_c, \qquad [5]$$

where M_c is defined in Eq. **1**.

DESCEND also computes standard errors and performs hypothesis tests on features of the underlying biological distribution, such as dispersion, nonzero fraction, and nonzero mean. See *Materials and Methods* for details.

Model Assessment and Validation

Technical noise model for UMI-based scRNA-seq experiments. For UMI-based scRNA-seq data, Kim et al. (14) gave an analytic argument for a Poisson error model, which we discuss and clarify in *SI Appendix*, *Mathematical Details of DESCEND*. Several studies (18, 34, 35) used spike-in sets and bulk RNA splitting experiments to explore the technical noise in scRNA-seq data, finding that a Poisson distribution for UMI-based counts is plausible, but raised the issue of overdispersion. While the analyses from these studies were insightful, we believe that they failed to account for the inevitable random variations across cells/samples in the spike-in input at low concentrations. We reexamined the spike-in data from nine UMI-based scRNA-seq datasets, covering seven scRNA-seq protocols (6, 7, 29, 36–40). All of the data except for those in ref. 29 have also been analyzed in ref. 40, which showed that capture efficiency varies substantially across cells within each experiment and across experimental protocols. We show that, after accounting for the cell-to-cell variation in efficiency, the technical noise of UMI counts is simply Poisson in most cases.

For External RNA controls Consortium (ERCC) spike-in "genes," the observed count for each gene in each cell depends on the number of input molecules and the technical noise. Due to the low spike-in concentration added to each cell, the number of input molecules for each spike-in is not fixed, but random with a certain target expectation. If we assume that the molecules in the spike-in dilution are randomly dispersed, then the number that results in each cell partition is Poisson with mean computable from the dilution ratios (see *SI Appendix*, *SI Text*, for more details). If the molecules in the spike-in dilution are not randomly dispersed, e.g., due to clumping, or if there are uncontrolled batch issues, then the input number of spike-in molecules for each cell would be overdispersed compared with the Poisson.

The key point here is that the input quantity of spike-in molecules is not fixed across cells, as assumed by current studies, but random with Poisson noise in the ideal case of perfect random dispersion with no batch variation. Such randomness in the input should not be counted as part of the technical noise of scRNA-seq experiments, as that is due to the spike-in process. Previous analyses of spike-in data have attributed this input-level variation to technical noise, thus inflating their estimates of technical noise dispersion.

To assess whether the $Y_{cg} \sim F_{cg}(\lambda_{cg}) = \text{Poisson}(\alpha_c \lambda_{cg})$ model fits the technical noise of each dataset, we did the following: DESCEND is applied to each spike-in gene in each dataset with this error model to obtain the underlying distribution of the input molecule counts. If this model is a good approximation to the true technical noise distribution of the scRNA-seq experiment, and if the spike-ins are ideal in the sense described above, then the DESCEND-recovered input molecule (λ_{cg}) distributions of the spike-in genes should be Poisson. Conversely, if the recovered distributions show zero inflation or overdispersion compared with the Poisson, then that may be due to either a misspecified technical noise model or unaccounted experimental factors in the spike-ins. Note that the default use of DESCEND does not require spike-ins; here, the spike-ins from these nine studies are simply used to assess whether the technical noise model assumed by DESCEND is appropriate.

Fig. 2*A* shows that the DESCEND-recovered distribution in all but one (37) of the nine UMI datasets has overdispersion $\theta < 0.015$ compared with the Poisson, where θ is defined in the variance-mean equation $\sigma^2 = \mu + \theta\mu^2$. The overdispersion

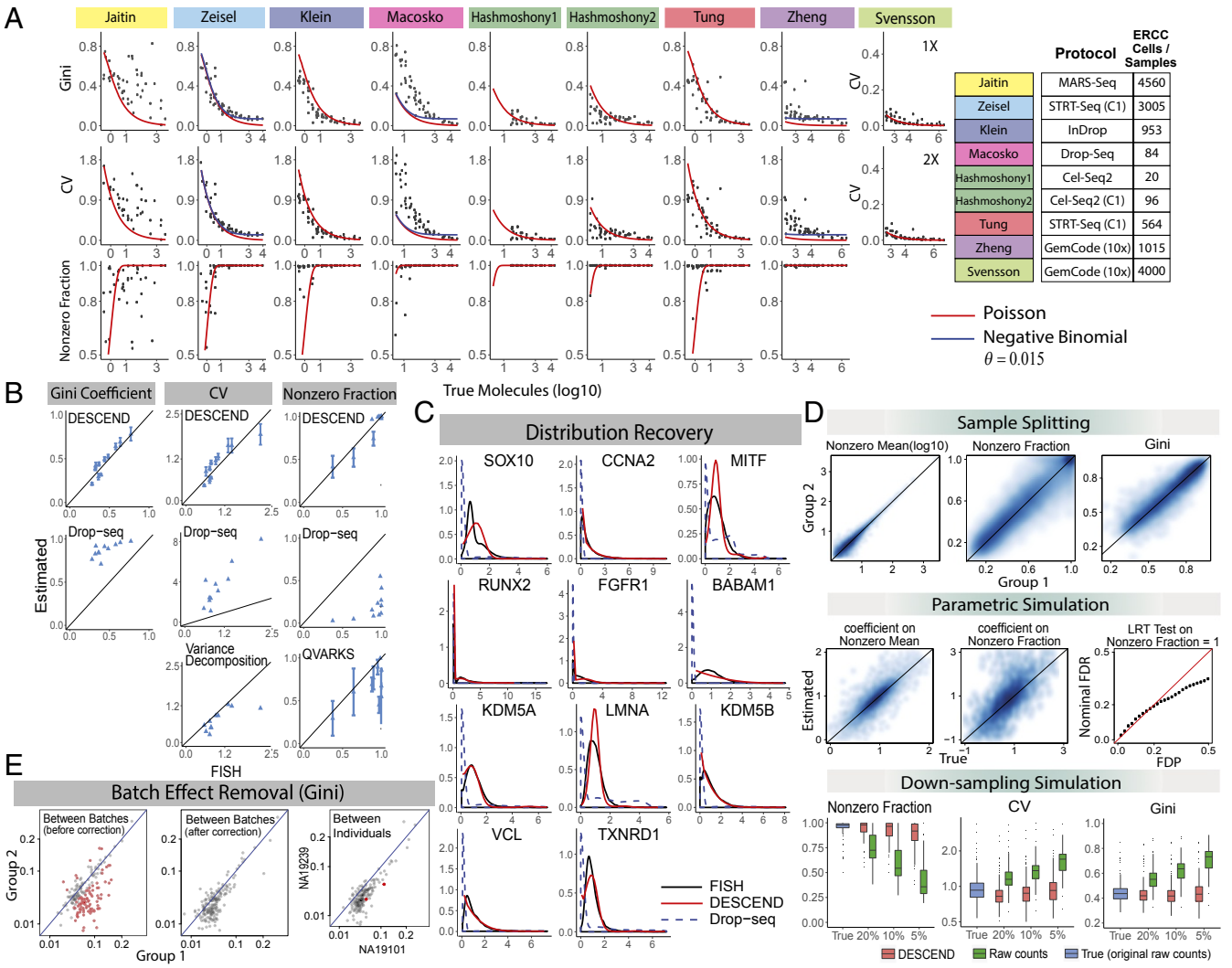


Fig. 2. Validation of DESCEND. (*A*) Noise model. The Poisson-alpha model is tested using nine different ERCC UMI datasets. Svensson et al. (40) include datasets at two different concentrations. The black dots are estimated quantities (Gini, CV and nonzero fraction) from the deconvolved distribution of each spike-in gene. The two solid curves show expected values of these quantities under the Poisson distribution (red) and negative binomial distribution with fixed $\theta = 0.015$ (blue). (*B*) RNA FISH. Gini, CV, and nonzero fraction of 11 genes are compared between RNA FISH and the DESCEND estimates from Drop-seq counts (13). Values computed directly from observed counts and by other methods are also included. (*C*) FISH distribution recovery. Relative gene expression distribution is compared among RNA FISH distribution, DESCEND, and the distribution of Drop-seq observed counts. (*D*) Simulations. For sample splitting, estimated quantities are compared between the two split groups. For the parametric simulation, coefficients of the covariate cell are compared with the true values. The false discovery proportion (FDP) is compared with nominal FDR. For the down-sampling simulation, boxplots of estimated and "true" (original raw counts) values across genes are compared. (*E*) Batch effect removal in Tung et al. (29). The DESCEND-estimated Gini for each gene is compared between two replicates before (*Left*) and after (*Center*) adding batches as covariates and between two individuals (*Right*) after batch adjustment. The red dots are the significantly differential genes (of Gini) when FDR in controlled at level 5%.

is effectively zero in six of the datasets and less than 0.015 in the other two, indicating that the technical noise model used by DESCEND well approximates the technical noise in the data. As discussed above, the misfit of the Poisson to the recovered distribution for (37) data can be due to either a wrong technical noise model or clumping in the spike-ins. Note that for ref. 37, the overdispersion is high for low input values, which is the reverse of that for typical RNA-seq experiments. This pattern of overdispersion can be explained by a clumping model on the input molecules (see *Materials and Methods* for discussion).

Evaluation of deconvolution accuracy using RNA FISH as gold standard. Next, we evaluate the accuracy of DESCEND on the data from ref. 13, where Drop-seq and RNA FISH are both applied to the same melanoma cell line. A total of 5,763 cells and 12,241 genes were kept for analysis from the Drop-seq experiment, with median 1,473 UMIs per cell. Of these genes, 24 were profiled using RNA FISH (*VCF* and *FOSL1* were removed from the original data; *SI Appendix*, *SI Text*). We further excluded genes with zero UMI count in more than 98% of the cells, resulting in 12 genes. Relative gene expression distributions were recovered by DESCEND and are compared with gene expression distributions observed by RNA FISH. Since distributions recovered by DESCEND reflect relative expression levels (i.e., concentrations), for comparability the expression of each gene in FISH was normalized by *GAPDH* (41).

Both CV and Gini coefficients recovered using DESCEND match well with corresponding values from RNA FISH (Fig. 2*B*) for all 11 genes (*GAPDH* excluded). In comparison, Gini and CV computed on the original Drop-seq counts, standardized by

library size N_c (1), show very poor correlation and substantial positive bias; this agrees with previous observations (6, 13). For CV, a variance decomposition approach adapted from ref. 6 (*SI Appendix*, *SI Text*) shows bias toward 0 compared with values calculated from RNA FISH. This analysis also shows that the 1-SD error bars given by DESCEND reasonably quantify the uncertainty of its estimates.

DESCEND provides reasonably accurate estimates of the nonzero fraction, despite the low sequencing depth of this dataset (Fig. 2*B*). In contrast, the naive estimate, derived from the proportion of nonzero raw counts for each gene, is grossly inflated due to the low sequencing depth and is not a reliable estimator of nonzero fraction. DESCEND outperforms a recent method QVARKS, which estimates the nonzero fraction (called "ON fraction" in ref. 25) using a Bayesian approach.

Finally, the shape of the relative gene expression distribution (Fig. 2*C*) given by DESCEND matches that from FISH. In comparison, the distribution of the library-size standardized observed counts is quite different from that of their FISH counterparts, showing severe zero inflation and increased skewness.

Assessment of estimation accuracy and test validity by simulations. We further evaluate the accuracy of DESCEND by in silico sample splitting and by parametric and down-sampling simulations. For all in silico evaluations, we start with the observed counts of the 820 oligodendrocyte cells from ref. 7, for which ERCC spike-ins are available to estimate cell-specific efficiencies. Details of each simulation are given in *SI Appendix*, *SI Text*.

First, in the sample-splitting experiment, the 820 cells are randomly split into two equal-sized groups. Since the data are split randomly, there should not be any real differences between groups. The agreement in DESCEND estimates of nonzero mean, nonzero fraction, and Gini coefficients between the two groups (Fig. 2*D*) indicates that the procedure has low variance and is robust to the randomness of observed counts.

The above experiment gives a model-free assessment of DESCEND estimation variance. To assess the estimation bias, we then performed a parametric simulation experiment where true gene expression counts were simulated from a lognormal distribution with cell size as covariate and where noise was simulated from our proposed noise model. True values of all involved parameters were set to be estimates from real data. We see that the estimation of covariate effects on the nonzero fraction/mean (Fig. 2*D*), for which there is no RNA FISH validation, is reliable. Nonzero fraction, CV, and the Gini also get accurate and unbiased estimates (*SI Appendix*, Fig. S1*A*). In addition, with the Benjamini–Hochberg procedure, DESCEND effectively controls the false discovery rate (FDR) in the test of whether the nonzero fraction is 1 (Fig. 2*D*).

Finally, we perform down-sampling simulations to assess how DESCEND performs under varying sequencing depth. The top 150 genes with highest total UMI count are selected and their UMI counts are treated as true expression levels. These values are then down-sampled at $\alpha = 20\%, 10\%$, and 5% efficiency levels. The nonzero fraction, CV, and Gini coefficients estimated by DESCEND are robust to change in efficiency level while their counterparts computed directly from raw counts are severely affected by such changes (Fig. 2*D* and *SI Appendix*, Fig. S1*A*).

There is, of course, an endless list of parameters for which evaluations can be performed. We have merely summarized here evaluations that are relevant for the case studies discussed later in this paper.

Batch effects can be removed in differential analysis by adding batch as covariate. Tung et al. (29) performed scRNA-seq on three human iPSC cell lines, with three technical replicates per cell line, and showed that there can be substantial variation between technical replicates. Ref. 29 further showed that simple ERCC spike-in adjustment and library size normalization cannot effectively remove this technical "batch effect." We apply DESCEND to these data to see whether using batch as a covariate can effectively remove technical differences between replicates.

Starting from the data of ref. 29, we created two groups of cells, each containing 150 cells obtained by pooling 50 cells randomly selected from each of the three individuals. Thus, the two groups of cells should have no biological differences. However, the replicates (batches) are manually chosen to preserve the technical batch effect between the two groups: The first group contains cells sampled from one replicate for each subject: NA19098 replicate 1, NA19101 replicate 2, and NA19239 replicate 1; the second group contains cells sampled from another replicate from each subject: NA19098 replicate 3, NA19101 replicate 1, and NA19239 replicate 2. With the two groups of cells constructed in this way, any detection made during differential testing must be a false positive due to the technical differences between replicates (batch effects).

DESCEND was applied to these data to test for differences in Gini coefficient and CV between the two groups (Fig. 2*E* and *SI Appendix*, Fig. S1*B*). Without consideration of batch, DESCEND indeed finds many (false positive) differences in both Gini and CV. However, with batch added as covariate in the DESCEND model, the dispersion estimates from the two groups are comparable, and no significant detections are made. The fact that spike-in–based normalization cannot effectively remove this batch effect, which is effectively removed by the DESCEND model, indicates that batch effects are gene specific. We also conducted differential dispersion analysis between two biologically different samples (the three replicates from NA19101 vs. the three replicates from NA19239), with batch as covariate, and found some significant changes in Gini. The fact that significant differences are found between biologically different samples, but not between biologically identical samples, suggests that DESCEND still has the power to detect biological signals while removing batch effect.

Case Studies

Cell-size effect on expression distribution and differential testing of nonzero fraction and mean. At the single-cell level, many genes are bursty, being inactive in some cells and active in other cells. In Eqs. **3** and **4**, we defined the nonzero fraction and nonzero mean to characterize this heterogeneity in the true underlying gene expression. Using DESCEND, we analyze the scRNA-seq data of mouse hippocampal region from Zeisel et al. (7), where the 3,005 cells were classified into seven major cell types. Our goal is to characterize the dependence of nonzero fraction and nonzero mean on cell size and to find genes that are differentially expressed in these parameters between different cell types, controlling for cell size. Recall that cell size is estimated as Eq. **5**.

First, consider the transcriptome-wide patterns of the DESCEND deconvolved nonzero fraction and nonzero mean, without adjusting for cell size; these were estimated for each gene in each cell type separately, with no added covariates (details in *SI Appendix*, *SI Text*). As shown in Fig. 3*A*, the deconvolved gene expression distributions for most genes have much larger nonzero fraction in the neuron cell types (CA1 pyramidal, S1 pyramidal, and interneurons) compared with the nonneuron cell types (astrocytes–ependymal, endothelial–mural, microglia, and oligodentrocytes), thus suggesting that more genes are in the "on" state in neurons. However, neurons are known to be larger cells, and, indeed, cell-size estimates are substantially larger for neurons compared with nonneurons in this dataset (*SI Appendix*, Fig. S2*A*). Can the global increase in nonzero fraction in neurons simply be attributed to increased cell size? To answer this question we need to first quantify how cell size affects the gene expression distribution.

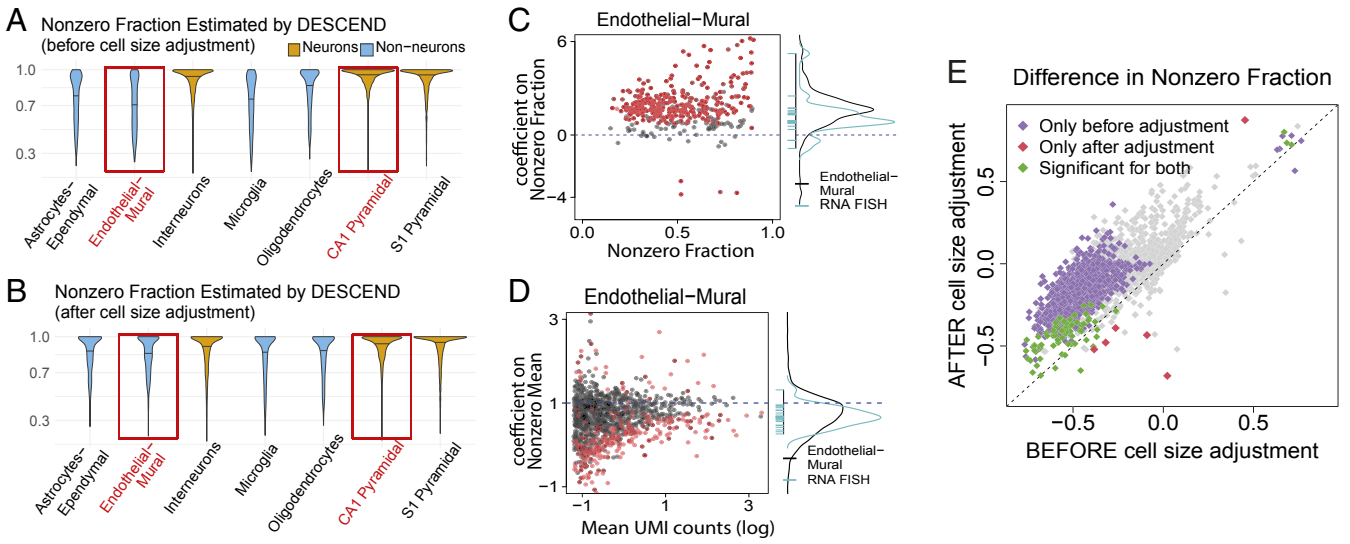


Fig. 3. Differential testing on nonzero fraction/mean as in Zeisel et al. (7). Violin plots of the estimated nonzero fractions are compared across cell types (*A*) before and (*B*) after adding cell size as a covariate. (*C*, *Left*) Estimated coefficients of cell size on nonzero fraction for genes whose nonzero fraction is significantly smaller than 1 and with estimated value less than 0.9 for the endothelial–mural cell population. (*C*, *Right*) Density of all of the dots in *C*, *Left* (black curve) aligned with the density curve of the coefficients of cell size on nonzero fraction for the RNA FISH data (blue). (*D*) Same as *C*, but for coefficients on nonzero mean and all of the genes. (*E*) Scatter plot for the difference of the estimated nonzero fraction between the endothelial–mural and CA1 pyramidal cells before and after cell-size adjustment. Significant genes are highlighted at FDR level 5%.

We applied DESCEND, with cell size as a covariate, to obtain the deconvolved cell-size–adjusted gene expression distribution for each gene in each of the seven cell types. The coefficients estimated by DESCEND allow us to assess, for each gene, whether its nonzero mean has superlinear, linear, or sublinear growth with cell size and whether its nonzero fraction increases, remains constant, or decreases with cell size. See statistical details in *Materials and Methods*. Taking the endothelial–mural cells as an example, we find that for most genes, nonzero fraction increases with cell size as the coefficients are positive (Fig. 3*C*). The mean trend across genes is that a doubling of cell size is associated with at least a doubling of the odds of observing at least one transcript. We also find that, globally, nonzero mean has a slightly sublinear dependence on cell size as many coefficients are below 1 (Fig. 3*D*). The sublinear dependence of nonzero mean on cell size is consistent with previous findings in ref. 41, which used RNA FISH to study a small set of genes and found their expression to have increased concentration in smaller cells, although the quantity measured in ref. 41 directly reflects transcription burst size. These relationships between expression distribution and cell size are consistent across all seven cell types in this study (*SI Appendix*, Fig. S2 *E* and *F*).

It is important to emphasize here that both cell size and true expression distribution are not directly observable quantities and that this relationship between cell size and nonzero mean/fraction is not a direct consequence of larger observed library size leading to larger change of having a nonzero count for a gene. To demonstrate further that the discovered relationships are biological, not technical, we conducted a parallel analysis of the RNA FISH data of ref. 13 (*SI Appendix*, *SI Text*). For the 23 genes (excluding GADPH) available in the RNA FISH data, we observe the same trends as above: The nonzero fraction increases with cell size, with a mean odds ratio of at least 2, and the nonzero mean increases sublinearly with cell size (Fig. 3 *C* and *D* and *SI Appendix*, Fig. S2*C*). The fact that this trend is observed using two different technologies and for eight different cell types (seven by scRNA-seq, melanoma cell line by RNA FISH) suggests that it may be a general relationship between cell size and single-cell gene expression distributions.

Fig. 3*B* shows the nonzero fractions across genes within each cell type, estimated by applying DESCEND with cell size as a covariate. After adjusting for differences in cell size, the transcriptome-wide patterns in nonzero fraction/mean are much more similar across cell types. This suggests that the increased nonzero fraction in neuron cells can mostly be attributed to cell-size differences. For example, compare two cell types: endothelial–mural and pyramidal CA1 cells. Before cell-size adjustment, 879 genes show significant decrease of nonzero fraction in pyramidal CA1 at FDR of 5% (Fig. 3*E* and *SI Appendix*, Fig. S2*B*); the results change substantially after cell-size adjustment, with only 84 significant genes, 78 of which were in the original set of 879 genes. This highlights the importance of cell-size adjustment in differential testing.

We also test for the change on nonzero mean between endothelial–mural and pyramidal CA1 cells. Across genes, the estimated change in nonzero fraction does not seem correlated with the estimated change in nonzero mean (Fig. 3*H*), indicating that, after accounting for cell size, the two types of change are unrelated. Differential testing results on nonzero mean and nonzero fraction, taken together, give a more detailed characterization of differential expression (*SI Appendix*, Fig. S3).

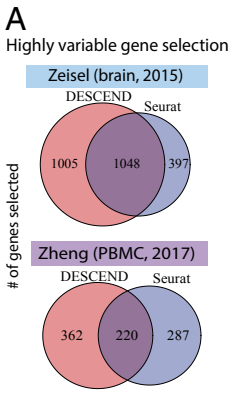
DESCEND improves the accuracy of cell type identification by a better selection of highly variable genes. One major step in cell type identification is the selection of highly variable genes (HVG) before applying any dimension reduction and clustering algorithms (6, 34, 37, 42). Filtering out genes with low variation reduces noise while keeping the genes that are more likely to be cell subpopulation markers. Current pipelines select HVGs mainly by computing dispersion measurements, such as CV and Fano factor, directly on the raw or library-normalized counts or by variance decomposition. However, as shown in Fig. 2*B*, these methods are not as accurate as DESCEND in estimating the true biological dispersion of the gene expression levels. Furthermore, compared with CV and Fano factor, the Gini coefficient is a more robust indicator of dispersion (see *Materials and Methods* for derivation), and we have shown that DESCEND allows accurate estimate of this indicator. Here, we examine whether

DESCEND-selected HVGs improve the accuracy of cell type identification when used with existing clustering algorithms.

We consider cell type identification in two datasets where somewhat reliable cell type labels are available. One consists of the 3,005 cells in ref. 7, which were classified into seven major cell types with the help of domain knowledge. The other is obtained from 10 purified cell populations derived from peripheral blood mononuclear cells (PBMC) in ref. 39, where 1,000 cells were sampled randomly from each cell type and combined to form a 10,000-cell dataset. Since the PBMC data consist mostly of immune cell subtypes ($CD4^+$ helper T cells, $CD4^+/CD25^+$ regulatory T cells, $CD4^+/CD45RA^+/CD24^-$ naive T cells, $CD4^+/CD45RO^+$ memory T cells, $CD8^+$ cytotoxic T cells, and $CD8^+/CD45RA^+$ naive cytotoxic T cells) which are well known to have similar transcriptomes, it is a more challenging test case. For both datasets, we treat the cell type labels given in their original papers as the gold standard.

There are many existing cell clustering methods for scRNA-seq. To focus on evaluating the effectiveness of the initial HVG selection step, we limit to Seurat, one of the most widely used algorithms, and compare clustering results of Seurat (Version 2.1) with a modified version of Seurat where the initial HVG selection step is replaced by DESCEND. Seurat selects HVGs by ranking the Fano factors of the normalized counts, yielding a list with only ∼50% overlap with the HVGs selected by DESCEND (Fig. 4*A*) for both datasets. To compare cell clustering accuracy, we use the adjusted Rand index, which ranges from 0, for a level of similarity expected by chance, to 1 for identical clusters (43). The number of clusters is set to the truth for both datasets and both pipelines. Fig. 4*B* shows that with DESCEND, Seurat achieves consistently better results than its original version. Seurat, like most other clustering algorithms, first performs dimension reduction using principal components analysis (PCA), and the number of principal components (PCs) affects downstream clustering. As seen in Fig. 4*B*, the accuracy boost obtained from DESCEND-based HVG selection is consistent across the number of PCs used for dimension reduction.

Dispersion analysis using the DESCEND-estimated Gini coefficient highlights early-stage differentiation markers for mES cells. We apply DESCEND to data from Klein et al. (6), where single cells were sampled from a differentiating mESC population before and at 2 d, 4 d, and 7 d after Leukemia inhibitory factor (LIF) withdrawal. In these data, while pluripotency markers and differentiation-related genes have changes in mean expression over time, due to complete transcriptome remodeling, almost all genes have significant change in mean expression even by day 2 (*SI Appendix*, Fig. S4*C*). Thus, change in mean is not an effective measure for identifying differentiation markers. Instead, we used DESCEND to test for change in expression dispersion across the early time points, under the rationale that early differentiation markers would exhibit high heterogeneity at this initial stage.



B

Cell Type Identification Accuracy (ARI)

# of PCs	Zeisel (K = 7) Seurat	Zeisel (K = 7) DESCEND + Seurat	Zheng (K = 10) Seurat	Zheng (K = 10) DESCEND + Seurat
10	0.671	0.838	0.519	0.556
15	0.674	0.887	0.520	0.571
20	0.826	0.866	0.520	0.574
25	0.702	0.864	0.522	0.572
30	0.732	0.869	0.519	0.574

Fig. 4. Selection of HVGs and cell type identification. (*A*) Venn diagram of the number of selected HVGs in Seurat and using DESCEND based on the Gini coefficient. (*B*) Comparison of cell type identification accuracy using Adjusted Rand Index (ARI) between the original Seurat and Seurat with the HVG selection step replaced by DESCEND.

First, consider expression dispersion as quantified by the Gini coefficients computed from the observed counts distribution, before deconvolution by DESCEND: For most genes, they are much higher at days 2 and 4, compared with days 0 and 7 (Fig. 5*A*). Although this pattern is consistent with our expectation that cells should exhibit higher heterogeneity during differentiation, compared with directly before or directly afterward, it is confounded by the substantially lower sequencing depth for the day 2 and 4 samples (*SI Appendix*, Fig. S4*A*). Are the higher Gini coefficients in days 2 and 4 due to the technical reason of lower sequencing depth or the biological reason of increased population heterogeneity? We applied DESCEND to compute Gini coefficients of the true expression distribution, thus removing the technical confounder of sequencing depth. DESCEND-estimated Gini coefficients confirm that Gini coefficients are indeed higher at days 2 and 4, compared with days 0 and 7 (Fig. 5*B*), as expected for this evolving population.

The differentiation of mES cells upon LIF withdrawal is a poorly characterized process. Ref. 6 observed that, whereas at day 7, almost 100% of cells have become epiblasts, at day 2 the proportion is below 10%. Thus, the cross-cell mean expression of known epiblast markers, such as *Krt8*, *Krt18*, *Tagln*, *Cald1*, *Tpm1*, and *Fxyd6*, does not show significant increase until day 4 (Fig. 5*C*), when almost half of the cells show complete transcriptome reprogramming (*SI Appendix*, Fig. S5*A*). In comparison, these known marker genes belong to a much smaller set of genes that show a dramatic increase in Gini coefficient between day 0 and day 2 (Fig. 5*C* and *SI Appendix*, Fig. S4*C*).

DESCEND allows the computation of SEs, and from these SEs we assessed the significance of change in Gini coefficient between day 0 and day 2. At an FDR threshold of 5%, 750 genes are significant for change in Gini coefficient. In comparison, more than 10,000 genes have significant, but very small changes, in mean between day 0 and day 2 using either DESeq2 or DESCEND, with the significance driven mostly by the much smaller SEs for mean estimation (*SI Appendix*, Fig. S4*C*). Of the 56 genes with significant change in Gini coefficient but not in mean computed by DESCEND, many are meaningful differentiation markers. For instance, the genes *Tagln*, *Anxa2*, *H19*, *Sparc*, and *Ccno* are listed in ref. 6 as marker genes for the cell types that appear during differentiation (*SI Appendix*, Fig. S4*D*). *Jun*, *Anxa3*, *Klf6*, *Fos*, and *Dusp4* can also be marker genes for the epiblast cells as these genes show significant increase in mean expression at the later stages (day 4 or day 7) (*SI Appendix*, Fig. S4*E*).

DESCEND also allows a more detailed characterization of the activity of pluripotency factors during differentiation (44, 45). As discussed in ref. 6, the expression levels of some pluripotency genes drop gradually (*Pou5f1*, *Dppa5a*, *Sox2*) while those of others drop rapidly (*Nanog*, *Zfp42*, *Klf4*) during differentiation. This is revealed by the trend of the DESCEND-estimated Gini coefficient (Fig. 5*E*) across time. The rapid drop-off genes react early during differentiation, and thus their Gini coefficients increase immediately at day 2 (*SI Appendix*, Fig. S5). In comparison, the gradual drop-off genes react late during differentiation and thus their Gini coefficients remain unchanged at day 2, starting to increase only at day 4. In contrast, this difference in early vs. late expression drop-off is not visible by mean expression due to heterogeneity between cells with regard to their differentiation timing. As a negative control, the cell-cycle marker gene *Ccnd3* has no significant change in DESCEND-estimated Gini

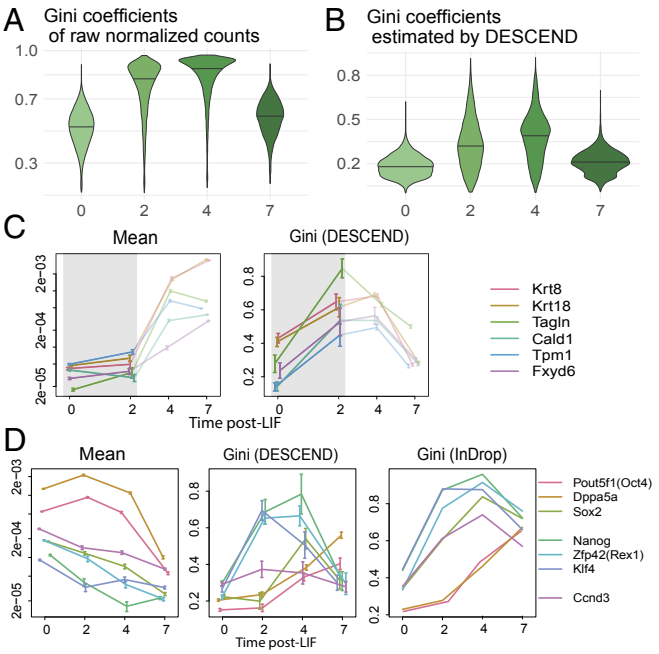


Fig. 5. Marker genes analysis using Gini as in Klein et al. (6). (*A* and *B*) Violin plots (with solid line indicating the 50% quantile) of Gini coefficients of raw normalized counts (*A*) and of the DESCEND-estimated Gini coefficients on each day (*B*). (*C*) Change of the mean relative expression and Gini coefficients for six epiblast marker genes across days. The colored error bars indicate 1 SE. (*D*) Change of the mean relative expression and Gini coefficients for pluripotency genes across days. For the Gini coefficients, one is estimated using DESCEND, and the other is calculated using raw normalized counts. The colored error bars indicate 1 SE.

coefficient during differentiation, agreeing with the fact that its expression heterogeneity across cells should be steady during differentiation.

Discussion

We have described DESCEND, a method for gene expression deconvolution for scRNA-seq. All results in this paper are based on a Poisson model for UMI counts. The deconvolution accuracy of DESCEND was extensively assessed through comparisons to population-matched RNA FISH data and through data splitting and simulation experiments. DESCEND's noise model allows it to remove known batch effects, as demonstrated on data from ref. 29.

DESCEND's formulation allows more complex noise models, which would be necessary for analysis of scRNA-seq data without UMIs where there is amplification bias and zero inflation beyond what could be captured by Poisson sampling (24, 46). But such models contain many more parameters, and the estimation of these parameters is nontrivial. Some aspects of the deconvolved distribution, such as Eqs. **3** and **4**, are highly sensitive to the noise model and the estimation quality of the technical parameters. More efforts are needed to develop robust error models for non-UMI scRNA-seq data.

Even in UMI-based scRNA-seq data, technical noise may have substantial overdispersion, and a negative binomial distribution may be more appropriate. DESCEND accepts the negative binomial distribution with a known overdispersion parameter θ. As shown in Fig. 2*A*, θ can be estimated using the spike-in genes as the square of the CV limit when the expected number of input molecules increases. The overdispersion may also be cell or gene specific, but a more realistic model with more parameters may not always lead to better analyses if those parameters cannot be estimated reliably from data. So far, we have settled on a simple model with at most one overdispersion parameter for UMI-based data.

Without covariate adjustment, DESCEND currently requires a few seconds for deconvolution of the distribution of a single gene with hundreds of cells and 10–20 s when there are thousands of cells. Adding covariates and performing likelihood-ratio tests increase the computation cost by roughly three or four times. Computation can be easily parallelized across genes.

Accuracy of DESCEND estimation increases with number of cells and with sequencing coverage. For example, for the Drop-seq data from ref. 13, although the average UMI count per cell is only around 1,500, the large number of cells (a few thousand) allows accurate DESCEND estimation. For the data from refs. 6 and 7, there are only a few hundred cells for each condition, but the high sequencing depth and the prefiltering allow good estimates.

Code Availability

The R package DESCEND is available on the Github repository https://github.com/jingshuw/descend. All source code and intermediate analysis results that were used to generate the figures in this paper are at repository https://github.com/jingshuw/DESCEND_manuscript_source_code.

Materials and Methods

Model. We introduce the model, estimation procedure, and inference framework here, but leave technical details and a full discussion to the accompanying mathematical supplement in *SI Appendix, Mathematical Details of DESCEND*.

The observed count Y_{cg} for gene g in cell c is modeled as a convolution of the true gene expression λ_{cg} and independent technical noise $F_{cg}(\cdot)$:

$$Y_{cg} \sim F_{cg}(\lambda_{cg}). \quad [6]$$

The default technical noise model in DESCEND is the Poisson-alpha model,

$$Y_{cg} \sim \text{Poisson}(\alpha_c \lambda_{cg}), \quad [7]$$

where α_c is a cell-specific scaling factor. There are two ways to set α_c, leading to two interpretations for λ_{cg}. First, if one wishes to recover the distribution of absolute expression, one would need spike-in data for each cell to compute the proportion of transcripts sequenced (2) and set α_c to this value. In this case, λ_{cg} represents true absolute expression count and α_c is the cell-specific efficiency constant. When spike-ins are not available, or when one wishes to recover the relative gene expression distribution, α_c can be set to the total UMI count for cell c in Eq. **1**. In this case, λ_{cg} represents the concentration of gene g in cell c. Thus, the interpretation of λ_{cg} depends on how we set α_c. From now on, we simply refer to λ_{cg} as "true expression."

DESCEND allows more complex technical noise models with gene-specific batch effects and Beta-binomial noise distribution. These extensions are described in *SI Appendix, Mathematical Details of DESCEND*.

Without covariates, our model assumes that true expression $\lambda_{cg} \sim H_g$ and our goal is to estimate H_g. We assume that H_g has a point mass at zero and a nonzero component belonging to an exponential family of distributions as in ref. 27. When cell-level covariates U_c are specified, we assume that they affect expression as follows:

$$\log(\lambda_{cg})|_{\lambda_{cg}>0} = U_c \beta_g + \epsilon_{cg}, \quad [8]$$

$$\text{logit}(\mathbb{P}[\lambda_{cg} = 0]) = U_c \tilde{\beta}_g + \tilde{\beta}_{0g}. \quad [9]$$

Thus, the covariate effect is quantified by the parameters β_g, $\tilde{\beta}_g$, and $\tilde{\beta}_{0g}$. Our goal is to conduct inference on these parameters as well as to deconvolve the "covariates-adjusted distribution." The nonzero part of the distribution is the same as that of $\exp(\epsilon_{cg})$ and the zero probability (denoted by p) of this distribution is the average of $\mathbb{P}[\lambda_{cg} = 0]$ across cells c. The covariates-adjusted nonzero fraction is meaningful only for differential testing, which is defined as $1/\left(\exp(\tilde{\beta}_{0g}) + 1\right)$, with U_c centered to have mean 0 across all cell populations.

Modeling the Nonzero Component. Let $h(\cdot)$ be the density of $\exp(\epsilon_{cg})$ or, in the absence of covariates, the density of $\lambda_{cg}|\lambda_{cg} > 0$. We assume

$$h(x) = \exp\{Q(x)^T\alpha - \phi(\alpha)\}, \quad [10]$$

where α is a vector of parameters and $\phi(\alpha)$ is the normalization factor. When $Q(x) = (\log x, x)$, then $h(x)$ is the Gamma density. When $Q(x) = (\log x, (\log x)^2)$, $h(x)$ is the lognormal density. Following ref. 27, to adapt to data, we set $Q(x)$ to the 5-degree natural cubic spline basis.

As in ref. 27, we discretize to simplify estimation. That is, we assume $\lambda \in \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ and let

$$\mathbb{P}[\lambda_{cg} = \lambda_j \mid \lambda_{cg} > 0] = \exp\{Q_j^T\alpha - \phi(\alpha)\}, \quad [11]$$

where $Q = (Q_1, \ldots, Q_m)^T$ is the 5-degree natural cubic spline base matrix. In DESCEND, the default setting is $m = 50$ and λ is chosen as equally spaced points between 0 and the $1 - a$ percentile of $\{Y_{cg}/\alpha_c,\ c = 1, 2, \cdots C\}$. See *SI Appendix, SI Text* for how a is chosen.

Penalized Maximum-Likelihood Estimation. Now we combine zero inflation and covariates adjustment with density h to get the likelihood of the observed data Y_{cg}, which we call f_c. It is not hard to show that the likelihood has the form

$$f_c(\tilde{\alpha}_g, \beta_g) = \boldsymbol{p}_c(\beta_g)^T \boldsymbol{h}_c(\tilde{\alpha}_g),$$

where $\boldsymbol{p}_c(\beta_g)$ incorporates the Poisson noise and the covariate effect on nonzero mean (Eq. **8**) (the formula for $\boldsymbol{p}_c$ is in ref. 27), and $\boldsymbol{h}_c(\tilde{\alpha}_g)$ is an adjustment of Eq. **11** to account for Eq. **9**,

$$\boldsymbol{h}_c(\tilde{\alpha}_g) = \exp\{Q_c^T\tilde{\alpha}_g - \tilde{\phi}_c(\tilde{\alpha}_g)\},$$

where

$$Q_c^T = \begin{pmatrix} 1 & U_c & 0 \\ 0 & 0 & Q^T \end{pmatrix}$$

is the cell-specific covariate-adjusted matrix. The first element of $\tilde{\alpha}_g$ is a rescaled $\tilde{\beta}_{0g}$ and the rest of $\tilde{\alpha}_g$ is $(\tilde{\beta}_g, \alpha)$.

Following ref. 27, we maximize a penalized log-likelihood to estimate the parameters $\beta_g, \tilde{\alpha}_g$. Suppressing g in our notation, let the log-likelihood for counts of gene g across cells be $l(\tilde{\alpha}, \beta) = \sum_c \log f_c(\tilde{\alpha}, \beta)$, and the penalized log-likelihood is $\tilde{l}(\tilde{\alpha}, \beta) = l(\tilde{\alpha}_g, \beta_g) - s(\tilde{\alpha})$, where $s(\tilde{\alpha}) = c_0\|\tilde{\alpha}\|_2$ is the penalty term. Let the Fisher information matrix of $\tilde{\alpha}$ be $I(\tilde{\alpha})$. Based on the suggestion in ref. 27, in DESCEND the tuning parameter c_0 is adaptively chosen such that the approximated ratio of artificial to genuine information $R(\widehat{\tilde{\alpha}}) = \mathrm{tr}\{\ddot{s}(\widehat{\tilde{\alpha}})\}/\mathrm{tr}\{\widehat{I}(\widehat{\tilde{\alpha}})\}$ is less than 1% to avoid overshrinkage but more than 0.05% to reduce overfitting.

Statistical Inference. Ref. 27 showed that second-order approximations provide useful inference on model parameters. By Taylor expansion of the log-likelihood around the true values of $\tilde{\alpha}$ and β, we have

$$0 = \left\{\dot{\tilde{l}}(\widehat{\tilde{\alpha}}, \widehat{\beta}) \approx \dot{\tilde{l}}(\tilde{\alpha}, \beta) + \ddot{\tilde{l}}(\tilde{\alpha}, \beta)\begin{pmatrix}\widehat{\tilde{\alpha}} \\ \widehat{\beta}\end{pmatrix} - \begin{pmatrix}\tilde{\alpha} \\ \beta\end{pmatrix}\right\}$$

from which one can calculate bias and SD of the estimates. For inference on functions of $\tilde{\alpha}, \beta$ (such as mean, nonzero fraction, etc.), we apply the Delta method.

Now consider differential testing between two populations of cells. Let θ_i ($i = 1, 2$) be the value of some model parameter, which is a function of $\tilde{\alpha}, \beta$, in population i. To test $H_0 : \theta_1 = \theta_2$ we compute a Z score,

$$Z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\widehat{\mathrm{MSE}}(\hat{\theta}_1) + \widehat{\mathrm{MSE}}(\hat{\theta}_2)}},$$

where $\widehat{\mathrm{MSE}}(\hat{\theta}_i) = \widehat{\mathrm{Bias}}^2(\hat{\theta}_i) + \widehat{\mathrm{SD}}^2(\hat{\theta}_i)$ is the estimated mean-squared error. Permutations of the cell labels give the null distribution for P-values computation.

DESCEND uses likelihood-ratio tests with the unpenalized likelihood $l(\tilde{\alpha}, \beta)$ to conduct tests on distribution parameters in a single population; e.g., $H_{01g} : \mathbb{P}[\lambda_{cg} \neq 0] = 1$, the test on nonzero fraction in the true distribution of gene g. Another example is the test of whether the cell-size covariate, in the first case study, has an effect on nonzero mean ($H_{02g} : \beta_g = 1$) or on nonzero fraction ($H_{03g} : \tilde{\beta}_g = 0$). These test statistics are approximately chi-square distributed.

Finding HVGs. We use quantile smooth regression (default quantile 0.5) to fit a smooth curve of the relationship between the mean of deconvolved distributions and Gini across genes, using the R package quantreg (47). The dispersion score of each gene is computed as the distance of the Gini from the curve, which is further normalized by its SE. We select HVGs as the genes whose normalized scores are larger than a threshold T (default value is 10).

Randomness of ERCC Spike-in Counts in scRNA-seq Experiments. Randomness of both the input counts and technical noise contributes to the randomness of ERCC spike-ins observed counts. First, the actual count of each spike-in gene in each cell deviates from its target count computed from dilution ratios and is approximately Poisson given the following three assumptions: (*i*) The molecules are distributed evenly in the dilution, (*ii*) each molecule moves independently, and (*iii*) the distribution of molecules in dilution does not change during the spike-in process. If any of these assumptions fail, the actual count landing in each cell would be overdispersed compared with Poisson. Such randomness is also observed empirically. For example, in ref. 7, 339 of the 3,005 cells have at least one ERCC gene (among the 38 spike-in genes with expected input count $\geq$5) whose observed UMI count is even larger than its expected input count, which would not have been possible if we assume the input count is fixed.

In the ERCC data of ref. 37, the variances of the DESCEND deconvolved distributions are roughly 10λ, where λ is the expected input count. This overdispersion can be explained by a clumping model: If 10 molecules on average move together in the dilution before they are added to cells, and if the clumps move independently, then the variance of the observed counts would be 10λ.

ACKNOWLEDGMENTS. J.W. is supported by NIH Grant 2 R01HG006137 and the Wharton Dean's Fund. M.L. is supported by NIH Grants R01GM108600, R01GM125301, and R01HL113147. N.R.Z. is supported by NIH Grant R01HG006137.

1. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK (2009) Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* 459:428–432.
2. Raj A, van Oudenaarden A (2009) Single-molecule approaches to stochastic gene expression. *Annu Rev Biophys* 38:255–270.
3. Tay S, et al. (2010) Single-cell NF-κB dynamics reveal digital activation and analog information processing in cells. *Nature* 466:267–271.
4. Loewer A, Lahav G (2011) We are all individuals: Causes and consequences of non-genetic heterogeneity in mammalian cells. *Curr Opin Genet Dev* 21:753–758.
5. Shalek AK, et al. (2014) Single cell RNA Seq reveals dynamic paracrine control of cellular variation. *Nature* 510:363–369.
6. Klein AM, et al. (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161:1187–1201.
7. Zeisel A, et al. (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347:1138–1142.
8. Shaffer SM, et al. (2017) Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* 546:431–435.
9. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: From theories to phenotypes. *Nat Rev Genet* 6:451–464.
10. Shalek AK, et al. (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498:236–240.
11. Jiang Y, Zhang NR, Li M (2017) Scale: Modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biol* 18:74.
12. Eberwine J, Sul JY, Bartfai T, Kim J (2014) The promise of single-cell sequencing. *Nat Methods* 11:25–27.
13. Torre E, et al. (2018) Rare cell detection by single-cell RNA sequencing as guided by single-molecule RNA FISH. *Cell Syst* 6:171–179.
14. Kim JK, Kolodziejczyk AA, Ilicic T, Teichmann SA, Marioni JC (2015) Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat Commun* 6:8687.
15. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA (2015) The technology and biology of single-cell RNA sequencing. *Mol Cell* 58:610–620.
16. Stegle O, Teichmann SA, Marioni JC (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* 16:133–145.
17. Kivioja T, et al. (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* 9:72–74.
18. Kim JK, Marioni JC (2013) Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol* 14:R7.
19. Finak G, et al. (2015) Mast: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 16:278.
20. Gu J, Du Q, Wang X, Yu P, Lin W (2015) Sphinx: Modeling transcriptional heterogeneity in single-cell RNA-seq. bioRxiv:027870.
21. Buettner F, et al. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* 33:155–160.
22. Kharchenko PV, Silberstein L, Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. *Nat Methods* 11:740–742.

23. Prabhakaran S, Azizi E, Carr A, Pe'er D (2016) Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. *Proceedings of The 33rd International Conference on Machine Learning, PMLR*, eds Balcan MF, Weinberger KQ (JMLR.org, New York), Vol 48, pp 1070–1079.
24. Vallejos CA, Marioni JC, Richardson S (2015) BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Comput Biol* 11:e1004333.
25. Narayanan M, Martins AJ, Tsang JS (2016) Robust inference of cell-to-cell expression variations from single- and k-cell profiling. *PLoS Comput Biol* 12: e1005016.
26. Korthauer KD, et al. (2016) A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol* 17:222.
27. Efron B (2016) Empirical Bayes deconvolution estimates. *Biometrika* 103:1–20.
28. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550.
29. Tung PY, et al. (2017) Batch effects and the effective design of single-cell gene expression studies. *Sci Rep* 7:39921.
30. McDavid A, et al. (2012) Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* 29:461–467.
31. Delmans M, Hemberg M (2016) Discrete distributional differential expression (D3E)–a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* 17:110.
32. Vu TN, et al. (2016) Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics* 32:2128–2135.
33. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4:e309.
34. Brennecke P, et al. (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* 10:1093–1095.
35. Grün D, Kester L, Van Oudenaarden A (2014) Validation of noise models for single-cell transcriptomics. *Nat Methods* 11:637–640.
36. Jaitin DA, et al. (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343:776–779.
37. Macosko EZ, et al. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161:1202–1214.
38. Hashimshony T, et al. (2016) CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-seq. *Genome Biol* 17:77.
39. Zheng GX, et al. (2017) Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 8:14049.
40. Svensson V, et al. (2017) Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* 14:381–387.
41. Padovan-Merhar O, et al. (2015) Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol Cell* 58:339–352.
42. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC (2017) Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nat Methods* 14:565–571.
43. Kiselev VY, et al. (2017) SC3: Consensus clustering of single-cell RNA-seq data. *Nat Methods* 14:483–486.
44. Chu LF, et al. (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol* 17:173.
45. Papatsenko D, Xu H, Ma'ayan A, Lemischka I (2013) Quantitative approaches to model pluripotency and differentiation in stem cells. *Stem Cells Handbook* (Humana Press, New York), pp 59–74.
46. Jia C, Kelly D, Kim J, Li M, Zhang N (2017) Accounting for technical noise in single-cell RNA sequencing analysis. bioRxiv:116939.
47. Koenker R (2017) *quantreg: Quantile Regression*, R package Version 5.34.