SEPARATING MEASUREMENT AND EXPRESSION MODELS CLARIFIES CONFUSION IN SINGLE CELL RNA-SEQ ANALYSIS

ABHISHEK SARKAR[1,†] AND MATTHEW STEPHENS[1,2,†]

[1] Department of Human Genetics, University of Chicago, Chicago, IL, USA
[2] Department of Statistics, University of Chicago, Chicago, IL, USA
[†] To whom correspondence should be addressed

Abstract. How to model and analyze scRNA-seq data has been the subject of considerable confusion and debate. The high proportion of zero counts in a typical scRNA-seq data matrix has garnered particular attention, and lead to widespread but inconsistent use of terminology such as "dropout" and "missing data." Here, we argue that much of this terminology is unhelpful and confusing, and outline simple ways of thinking about models for scRNA-seq data that can help avoid this confusion. The key ideas are: (1) observed scRNA-seq counts reflect both the actual expression level of each gene in each cell and the measurement process, and it is important for models to explicitly distinguish contributions from these two distinct factors; and (2) the measurement process can be adequately described by a simple Poisson model, a claim for which we provide both theoretical and empirical support. We show how these ideas lead to a simple, flexible statistical framework which encompasses a number of commonly used models and analysis methods, and how this framework makes explicit their different assumptions and helps interpret their results. We also illustrate how explicitly separating models for expression and measurement can help address questions of biological interest, such as whether mRNA expression levels are multi-modal among cells.

1. Introduction

The development of single cell RNA sequencing (scRNA-seq) technologies has greatly facilitated investigation of important biological questions that were previously difficult or impossible to study. However, how to model and analyze scRNA-seq data has been the subject of considerable confusion and debate. In particular, the high proportion of zero counts in a typical scRNA-seq data matrix has garnered special attention, and has lead to widespread but inconsistent use of terminology such as "dropout" and "missing data." In this paper, we argue that much of this terminology is confusing and unnecessary, and we outline simple ways of thinking and talking about scRNA-seq data that can help reduce confusion.

The first key idea informing our perspective is that observed scRNA-seq counts reflect two distinct factors: the variation in actual expression levels among cells, and the imperfect measurement process. Therefore, models for observed scRNA-seq counts, which we will call *observation models*, are obtained by specifying: (1) an *expression model* that describes how the true expression values vary among cells/genes, and (2) a *measurement model* that describes how observed counts deviate from the true underlying expression values. Carefully distinguishing between observation, expression, and measurement models is important both for avoiding confusion and for performing useful analyses. Indeed, a key goal of most RNA-seq analyses is to draw inferences about true expression values from observed counts, and this is impossible without explicit consideration of how the observed counts

are related to the expression values through a measurement process. Moreover, making measurement and expression models explicit can help clarify the underlying assumptions and aid interpretation of results such as parameter estimates.

The second key idea informing our perspective is that the measurement process for scRNA-seq counts can be captured using a very simple measurement model, based on a Poisson distribution. We give both theoretical arguments and empirical support for this model, and explain how a Poisson measurement model can capture the abundance of zeros in scRNA-seq data, without any need for special terminology or special treatment.

Both of these key ideas are simple, and neither is new. Indeed, the importance of distinguishing among sources of variation contributing to observed data and explicit consideration of the measurement process has a long scientific history[1], and the use of Poisson models for measurement variation in RNA-seq dates back to early work in this field[2,3]. However, many papers on scRNA-seq analysis do not incorporate these simple ideas. Most papers that describe models for scRNA-seq data focus exclusively on observation models, with measurement and expression models left implicit (ref.[4] is a notable exception). This ambiguity is especially problematic when the models include components said to capture, e.g., zero-inflation, without clearly indicating whether these are part of the measurement model, the expression model, or both, making it difficult to interpret their results. Here, we show how many different observation models which have been proposed for scRNA-seq data can be interpreted within a general framework which combines a Poisson measurement model with different expression models. This interpretation clarifies the assumptions about variation in true gene expression levels that underlie commonly used observation models for scRNA-seq data.

These simple ideas can also help frame and address questions of biological interest. For example, the question of whether gene expression patterns are multi-modal among cells is a question about the expression model, and not about the observation model. We investigate this question empirically in diverse datasets, and find that data are often consistent with surprisingly simple expression models. Specifically, a Gamma distribution often suffices to capture variation in expression levels among cells.

2. A call to simplify terminology

One major source of confusion in scRNA-seq analysis is the widespread but inconsistent use of terminology, especially "dropout," "missing data," "imputation," and "zero-inflation." Choice and use of terminology has many important consequences: it affects the way that researchers in the field think, the way they develop and apply methods, and the way they interpret results. We therefore begin by reviewing these terminologies, and explain why in many cases we view their continued use as unhelpful.

The term "dropout" has become commonly used in connection with the zeros in scRNA-seq data[5–8]. Historically, dropout referred to a particular failure mode of PCR, *allelic dropout*, in which specific primers would fail to amplify sequences containing a particular allele of a polymorphism, leading to genotyping errors for heterozygous individuals[9,10]. In scRNA-seq, the term "dropout" was introduced to describe a supposed failure mode that might cause a gene to appear highly expressed in one cell and not detected in another[5]. Although the precise source of such a failure mode was not specified, this usage seems intended to refer to some aspect of the RNA-seq measurement process. Since then, use of the term "dropout" has spread rapidly, but its meaning varies among papers and presentations (and sometimes even within papers and presentations!). For example, it is sometimes used simply as a synonym for the observed zeros, sometimes used to refer to the unknown subset

of zeros that correspond to genes that are expressed but undetected, and sometimes used to describe the fact that not all molecules present in the original sample are observed (a fact that affects counts at all genes, not only the zeros). Such wide variation in usage naturally leads to confusion and disagreement about how these phenomena should be modeled. For this reason, we argue the term "dropout" should be avoided in the context of scRNA-seq data. Instead, zero observations should simply be referred to as zeros, and measurement models should focus on the simple fact that not all molecules in the original sample are observed, with zeros being just one byproduct of this fact.

The term "missing data" is also commonly used in connection with zeros in scRNA-seq[7,11]. This terminology is misleading because the zeros are not missing data, at least not in the way this term is conventionally used in statistics. To illustrate this point, consider a statistical survey where a researcher counts the number of cars arriving at a particular intersection each minute. If no cars arrive in a particular minute then the observation would be recorded as zero; however this is not a "missing" observation. In contrast, if the researcher takes a lunch break, and does not record the arrivals for a particular hour, this would lead to sixty genuinely missing observations. The zeros in an scRNA-seq data experiment are more like the former than the latter: there is no analogue of a lunch break in the scRNA-seq measurement process. (Of course, in RNA-seq data the observed zeros are noisy, in that they do not necessarily imply that there were zero molecules present in the original cell; however, the same is true of the ones, and indeed all of the counts, and this can and should be accounted for in the measurement model.)

Unfortunately, inappropriate use of the term "missing data" has lead researchers to apply inappropriate statistical methods for dealing with missing data to scRNA-seq data. For example, in other statistical applications it is common to attempt to "impute" (fill in) missing values, and so viewing the zeros in scRNA-seq as missing has led to many papers describing methods to "impute" the zeros[7,12–16]. However, since the zeros in RNA-seq are not actually missing data, the meaning of these imputed values is unclear. The meaning may vary from method to method, but in no case do they represent the actual number of molecules of a gene present in the original cell. Thus, like "dropout", the term "missing data" should be avoided in scRNA-seq data, and methods that "impute" (only) the zeros should be avoided as they lack principle. Instead, every observation in scRNA-seq, including the zeros, should be treated as a noisy observation of an underlying true expression level, and inference should focus on clearly-defined tasks such as estimating these true expression levels.

Finally, the term "zero-inflated" is another common source of confusion and debate. In statistics, "zero-inflated" is used to describe any model for count data that is obtained from a simpler model by increasing the proportion of zeros. For example, "zero-inflated Poisson" refers to a distribution obtained by taking a standard Poisson distribution and then increasing the proportion of zeros. In scRNA-seq applications, the use of zero-inflated models is an understandable reaction to the very high proportion of zeros in the counts matrix. However, recent work suggests that zero-inflated models may be completely unnecessary; indeed ref.[17] entitles their piece "Droplet scRNA-seq data is not zero-inflated." Here, we take a slightly different perspective that emphasizes the distinction between measurement and observation models. Specifically, we argue that there is no convincing evidence to support the use of zero-inflated models for the measurement process. However, zero-inflated models could be appropriate for the observation model in some cases, depending on the actual variation in underlying expression among cells (i.e., the expression model). This perspective makes clear that the need for zero-inflated models may vary from data set to data set and from gene to gene, because it depends on variation in actual expression levels,

not on technical factors. This said, our empirical analyses (Section 4) suggest that simple observation models without zero-inflation suffice for many data sets.

These terminological issues all stem from one major misconception in scRNA-seq: the idea that the measurement process involves some distinct zero-producing technical mechanism that must be explicitly modeled. Indeed, some published observation models include a component that randomly creates zero observations irrespective of the actual underlying expression in the cells[14,16,18]. Such a mechanism would require a coordinated or systematic effect that conspires to miss all molecules from a particular gene in a particular cell, and there are no convincing theoretical arguments or empirical evidence supporting this idea. Thus, we argue that measurement models for scRNA-seq should begin from a simple assumption: the measurement process operates independently on each molecule in the cell, and each such molecule is either observed or not observed. This assumption leads to simple but flexible observation models, and simple explanations for the large proportion of zeros in scRNA-seq data. In this paper, we provide theoretical arguments and empirical evidence supporting this high-level view.

3. Modeling scRNA-seq data

Observed scRNA-seq counts reflect two distinct factors: the true expression levels of each gene in each cell, and the measurement process. In this section, we describe models for both measurement and expression, and how they combine to yield models for observed counts. We focus on scRNA-seq data that are generated using Unique Molecular Identifiers (UMIs), which ensure that each molecule in the cell is counted at most once[19]. Most contemporary scRNA-seq experiments use UMIs, which substantially reduce or eliminate several sources of unwanted variation including differences in gene lengths and PCR amplification efficiencies. The measurement models described here ignore these sources of unwanted variation, and so may be unsuited to data generated without UMIs.

Introducing notation, consider using scRNA-seq to measure gene expression in n single cells. Think of each cell as being a pool of mRNA molecules, let m_{ij} denote the true (but unknown and unobserved) number of mRNA molecules present in cell i from gene j ($i = 1, \dots, n; j = 1, \dots, p$), and let m_{i+} denote the total number of molecules in cell i. We refer to m_{ij} as the *absolute expression* of gene j in cell i and $\lambda_{ij} \triangleq m_{ij}/m_{i+}$ as the *relative expression* of gene j in cell i. We use $\mathbf{M} = [m_{ij}]$ and $\mathbf{\Lambda} = [\lambda_{ij}]$ to denote the matrices of expression values. Let $\mathbf{X} = [x_{ij}]$ denote the observed count matrix, where x_{ij} denotes the number of distinct molecules from gene j observed in cell i, and let x_{i+} denote the total number of molecules observed in cell i.

With this notation in hand, we can be more precise about measurement, expression, and observation models. A measurement model is a model that connects the observed counts $\mathbf{X}$ to the expression values $\mathbf{\Lambda}$, by specifying the conditional distribution $p(\mathbf{X} \mid \mathbf{\Lambda})$. An expression model is a model for the expression values $\mathbf{\Lambda}$, $p(\mathbf{\Lambda})$. Together, these two models determine the observation model, which is a model for the observed counts $p(\mathbf{X})$.

3.1. Modeling scRNA-seq measurement.

To derive a measurement model for scRNA-seq we assume: (1) in each cell, whether or not each molecule is observed is independent of whether or not each other molecule is observed, and (2) only a small proportion of all molecules present in each cell are observed. These assumptions are plausible because: (1) the measurement process operates at the molecular level, and (2) it is estimated that only 10–20% of all molecules present are actually captured and sequenced[8,20]. These assumptions

together imply the following Poisson measurement model (Appendix A)

$$x_{ij} \mid x_{i+}, \lambda_{ij} \sim \text{Poisson}(x_{i+}\lambda_{ij}). \tag{1}$$

Biases in the capture (library preparation) and/or sequencing processes that make some molecules more likely to be observed than others can also be accommodated within this Poisson measurement model. This measurement model is supported by theoretical arguments previously made for bulk RNA-seq data[3] and more recently in scRNA-seq data[21]. It is also supported by empirical results in early analyses of bulk RNA-seq data[2], more recent analyses of control scRNA-seq data[21,22], and our own analysis (Section 4).

Although the Poisson measurement model is essentially the same as that used for bulk RNA-seq data, it can nonetheless account for the fact that there are many more zeros in scRNA-seq data than bulk RNA-seq data for two simple reasons. First, the total number of molecules observed x_{i+} is typically orders of magnitude smaller for single cells than for bulk samples. Second, it is much more common that λ_{ij} will be very small (or even zero) for a single cell than for a bulk sample, because expression levels in bulk samples are averages of expression levels in many single cells. Intuitively, averaging reduces variance, moderating the occurrence of both very small and very large values. These two facts imply that the rate parameter $x_{i+}\lambda_{ij}$ is often much smaller in scRNA-seq data than in bulk RNA-seq data, explaining the much higher proportion of zero observations.

Despite its simplicity, the Poisson measurement model also captures aspects of scRNA-seq data that have previously been referred to using terms such as "dropout," "missing data," and "technical zeros." It captures the fact that the count matrix does not include every molecule that was present in every cell; indeed, this fact is a fundamental assumption in its derivation. It also captures the fact that x_{ij} may be observed to be zero even when m_{ij} (and hence λ_{ij}) is non-zero. However, the Poisson model captures both these features without introducing a distinct zero-generating mechanism. The zeros, like other observations, are simply imperfect measurements of the actual underlying expression, with no need for special terminology or special treatment.

Under the Poisson measurement model, observing $x_{ij} = 0$ is different from x_{ij} being missing. If x_{ij} were missing, then it would imply no information about the expression level λ_{ij}. But under the Poisson model, observing $x_{ij} = 0$ does provide such information, namely that λ_{ij} is unlikely to be large. Therefore, it makes no sense to treat the zero counts as missing data, or to attempt to impute their "true values" using methods developed for dealing with missing data. However, several existing methods[18,23] have instead relied on the idea that x_{ij} are noisy observations of the true underlying expression values, which we argue is a more useful starting point for models and analysis methods.

3.2. **Modeling gene expression.** At first sight, it may seem that the Poisson measurement model contradicts previous literature. Many papers have demonstrated that a Poisson distribution is inadequate to explain observed RNA-seq data at certain genes, and have adopted more general models such as negative binomial (NB) and zero-inflated negative binomial (ZINB) models for observed counts. However, in fact there is no contradiction here because the Poisson model here is a measurement model, whereas the NB and ZINB models are both used as observation models. Both NB and ZINB observation models can be derived by combining the Poisson measurement model with certain expression models. More generally, many existing models and methods for scRNA-seq data can be understood as instances of a general framework that combines the Poisson measurement model with specific expression models.

Expression model	Observation model	Method
Point mass (no variation)	Poisson	Analytic
Gamma	Negative Binomial	MASS, edgeR[26], DESeq2[27], BASICS[28]
Point-Gamma	Zero-inflated Negative Binomial	PSCL
Unimodal (non-parametric)	Unimodal	ashr[25,29]
Point-exponential family	Flexible	DESCEND[22]
Fully non-parametric[30]	Flexible	ashr

TABLE 1. **Single gene models for scRNA-seq data.** Different expression models, when combined with the Poisson measurement model, yield different observation models. Method indicates previously published methods and software packages that use these observation models to analyze data.

To make this idea precise, first consider modeling expression at a single gene j. An expression model for a single gene involves specifying a distribution g_j for the expression values $\lambda_{1j}, \dots, \lambda_{nj}$

$$\lambda_{ij} \sim g_j(\cdot). \tag{2}$$

One simple (and convenient) choice for the expression model is to assume that g_j is a Gamma distribution. Combining this Gamma expression model with the Poisson measurement model (1) yields an observation model that is NB, meaning that the counts $x_{1j}, \dots, x_{nj}$ follow an NB distribution[24].

Assuming a Gamma expression model is simple, but also somewhat restrictive. A more flexible, but still tractable, observation model can be obtained by replacing the Gamma expression model with a *point-Gamma* model, wherein some proportion of the expression values are exactly zero, while the other (non-zero) values follow a Gamma distribution. The point-Gamma expression model includes the Gamma model as a special case (when the proportion of exact zeros is 0), but can additionally capture one aspect of bi-modality of expression levels among cells, where some cells have no expression and others have non-zero expression. Combining a point-Gamma expression model with the Poisson measurement model yields an observation model that is ZINB.

The Gamma and point-Gamma model are both examples of parametric expression models; however, it is also possible to use non-parametric expression models[4,25]. These expression models for single genes, and some published methods implementing statistical inference for the corresponding observation models, are given in Table 1 (see Appendix B for details).

Similar ideas apply to expression models for multiple genes, although inevitably more complex. A multi-gene expression model simultaneously describes correlations among expression values at different genes across cells, and different cells across genes. A common and powerful approach to describe these correlations is to use low rank models, also known as factor models. The simplest such expression model assumes that the matrix $\mathbf{\Lambda}$ is low rank, meaning it can be factorized into two smaller rank k matrices $\mathbf{\Lambda} = \mathbf{L}\mathbf{F}'$, where the *loadings* matrix $\mathbf{L}$ is $n \times k$ and the *factors* matrix $\mathbf{F}$ is $p \times k$. This model requires a constraint that entries of $\mathbf{L}$ and $\mathbf{F}$ are non-negative, because the true expression values $\mathbf{\Lambda}$ must be non-negative. Combining this low rank expression model with the Poisson measurement model (1) yields the Poisson non-negative matrix factorization (NMF) observation model[31,32]. This model was initially proposed as minimizing a generalized KL divergence[33], which is equivalent to maximizing the Poisson log likelihood[34].

Link function	Noise distribution	Method
Identity	None	NMF[33], scHPF[37]
Identity	Gamma	NBMF[35]
log	None	GLM-PCA[21]
log	Gamma	scNBMF[38], GLM-PCA[21]
log	Point-Gamma	ZINB-WaVE[39]
Neural network	Point-Gamma	scVI[15], DCA[23]

TABLE 2. **Multi-gene models for scRNA-seq data.** Multi-gene models partition variation in true expression into structured and stochastic components. The link function describes a transformation, and the noise distribution indicates an assumption about the stochastic component. Method indicates previously published methods and software packages that use the corresponding observation model to analyze data.

Other existing methods can be viewed as a generalization of this idea in which $\mathbf{\Lambda}$ is assumed to be a low rank matrix perturbed by noise. For example, assuming an expression model $\lambda_{ij} = \mu_{ij} u_{ij}$, where $\boldsymbol{\mu}$ is a low rank matrix and the perturbations u_{ij} follow a Gamma distribution with mean one, leads to the negative binomial matrix factorization (NBMF) observation model[35]. This model has an appealing biological interpretation: the low rank matrix $\boldsymbol{\mu}$ represents the structure of expression variation among cells/genes and the perturbations u_{ij} represent stochastic deviations from this structure. Then, one might think of $\boldsymbol{\mu}$ as representing variation due to e.g., different cell types/states and u_{ij} as representing expression noise[36].

An even more general form of this idea is obtained by incorporating a transformation (e.g., log) into the low rank modeling framework. Many methods (including those discussed above) are obtained by combining the Poisson measurement model with an expression model of the form:

$$\begin{aligned} \lambda_{ij} &= \mu_{ij} u_{ij} \\ \mu_{ij} &= h^{-1}((\mathbf{LF}')_{ij}) \\ u_{ij} &\sim p_u(u_{ij}) \end{aligned} \tag{3}$$

where h denotes a *link function*, or transformation, and p_u denotes the assumed distribution of the perturbations. Finally, the most general form of this idea is obtained by assuming $\mu_{i\cdot} = h(l_{i\cdot})$, where h is a function (sometimes represented by a neural network) to be learned from the data. These multi-gene expression models, and some published methods implementing statistical inference for the corresponding observation models, are given in Table 2 (see Appendix C for details).

To summarize, many existing observation models for scRNA-seq data can be derived from combining the Poisson measurement model (1) with specific expression models of the form (2) or (3). This general statistical framework clarifies several sources of confusion in scRNA-seq analysis. First, it emphasizes that which observation models are most appropriate for scRNA-seq data may vary from data set to data set, and even from gene to gene within a single data set, because variation in expression is a biological phenomenon that will vary among data sets and genes. For example, if data are collected on a set of relatively homogeneous cells, then most genes might show relatively little variation in expression and for most genes it may suffice to model expression using a simple distribution. On the other

Box 1 Inference in the ZINB observation model

In our framework, the *Zero-inflated Negative Binomial* (ZINB) model for observations $x_{1j}, \dots, x_{nj}$ is written

$$x_{ij} \mid x_{i+}, \lambda_{ij} \sim \text{Poisson}(x_{i+}\lambda_{ij})$$
$$\lambda_{ij} \sim g_j(\cdot) = \pi_j \delta_0(\cdot) + (1 - \pi_j)\,\text{Gamma}(\phi_j^{-1}, \mu_j^{-1}\phi_j^{-1}).$$

where g_j is a *point-Gamma distribution*, which is a mixture of a point mass on zero (denoted δ_0) and a Gamma distribution (parametrized by shape and rate). Here, we consider two analysis tasks: (1) estimating g_j (i.e., estimating π_j, μ_j, ϕ_j), and (2) estimating λ_{ij}. Task (1) can be accomplished by maximizing the marginal likelihood, which has an analytic form but requires numerical optimization. From an estimate $\hat{g}_j$, one can already estimate many useful quantities such as the mean and variance of gene expression[40]

$$E[\lambda_{ij} \mid \hat{g}_j] = (1 - \hat{\pi}_j)\hat{\mu}_j$$
$$V[\lambda_{ij} \mid \hat{g}_j] = (1 - \hat{\pi}_j)\hat{\mu}_j^2\hat{\phi}_j + \hat{\pi}_j(1 - \hat{\pi}_j)\hat{\mu}_j^2.$$

Task (2) can be accomplished by estimating the conditional distribution of λ_{ij} given the observed data and $\hat{g}_j$

$$\lambda_{ij} \mid x_{ij}, x_{i+}, \hat{g}_j \sim \hat{\pi}_j \delta_0(\cdot) + (1 - \hat{\pi}_j)\,\text{Gamma}(x_{ij} + \hat{\phi}_j^{-1}, x_{i+} + \hat{\phi}_j^{-1}\hat{\mu}_j^{-1}).$$

If one interprets g_j as a prior, then this procedure is empirical Bayes, and the mean of the conditional distribution above is the posterior mean estimate of λ_{ij} given the observed data

$$E[\lambda_{ij} \mid x_{ij}, x_{i+}, \hat{g}_j] = (1 - \hat{\pi}_j)\frac{x_{ij} + \hat{\phi}_j^{-1}}{x_{i+} + \hat{\phi}_j^{-1}\hat{\mu}_j^{-1}}.$$

hand, if the data contain many cell types, then more complex multi-modal distributions could perhaps be required. We study this question empirically in Section 4.

Second, this framework provides a new interpretation for what it would mean to find, for example, that a ZINB model fit the counts (at a particular gene in a particular data set) better than an NB model: it would imply that the expression levels at this gene are better modeled by a point-Gamma distribution than by a simple Gamma distribution. Notably, this is a conclusion about the underlying expression values, not a conclusion about the measurement process. This interpretation contrasts with the usual way that the ZINB model is interpreted, in which the zero-inflation in the ZINB observation model captures some technical mechanism in the measurement process which could generate additional zeros, and the difference between a NB versus ZINB observation model relates to that technical mechanism. We examine this question of interpretation further in Appendix D.

Finally, this framework provides a statistically rigorous approach to infer actual expression levels from observed counts. Having assumed a Poisson measurement model, one can estimate the underlying expression model (and therefore features such as the mean or variance of true gene expression), as well as the true expression values themselves, from the observed counts directly. Importantly, these estimates are generally neither simple functions of the observed counts, nor functions of simple transformations (e.g., log) of the observed counts. We illustrate this procedure for the point-Gamma expression model (ZINB observation model) in Box 1.

4. Empirical examples

4.1. **Single gene models.** There is considerable debate about whether scRNA-seq data are adequately modeled by an NB observation model, or if it is necessary to use a ZINB (or even more complex) observation model. Using the framework and results from the previous

Dataset	Protocol	Number of cells	Number of genes	Source
Control 1	10X Chromium v2	2,000	88	[17]
Control 2	10X Chromium v2	2,000	88	[17]
Control	Drop-Seq	84	81	[17]
Control	GemCode	1,015	91	[17]
Control	InDrops	953	103	[17]
(Sorted) T cells	10X Chromium v2	10,209	6,530	[41]
(Sorted) B cells	10X Chromium v2	10,085	6,417	[41]
iPSC	Fluidigm C1	5,597	9,957	[40]
T cell/B cell mix[a]	10X Chromium v2	20,294	6,647	[41]
Cytotoxic T/Naive T mix[a]	10X Chromium v2	20,688	6,246	[41]
PBMCs	10X Chromium v2	68,579	6,502	[41]

TABLE 3. **Data sets analyzed.** Number of cells passing the previously reported QC filters and number of genes with non-zero observations in at least 1% of cells[41], or passing the previously reported QC filters[40] (a) Mixture data sets are generated *in silico* by concatenating the data and then applying QC filters.

section, these questions about the observation model can now be translated into questions about the expression model: is a Gamma expression model adequate, or is it necessary to use a more complex expression model? Importantly, this formulation makes clear that the answer may vary among genes and data sets because it depends on the actual variation in gene expression, not on technical factors.

To assess whether a Gamma expression model (or any other expression model) is adequate to capture variation in true gene expression levels at each gene, we applied a simple goodness-of-fit test[40]. This approach assesses the fit of an expression model by computing the value of the corresponding cumulative distribution function F for the observation model at the observed data points, and applying the Kolmogorov–Smirnov test to assess whether those values depart from a uniform distribution (Appendix B). If the values depart from a uniform distribution, then there is a discrepancy between the observations and the fitted F (Figure 1A). We applied this test to a variety of scRNA-seq data sets, comprising both control data sets in which synthetic mRNA molecules are directly added ("spiked in") to droplets at known concentrations, captured, and then sequenced[17], and biological data sets spanning a range of settings including: homogeneous collections of sorted cells, *a priori* homogeneous cell lines, heterogeneous tissues, and *in silico* mixtures of sorted cells designed to produce heterogeneous expression patterns (Table 3).

In control data sets, we found many spike-in genes depart from a point mass expression model (i.e., a model of no variation), consistent with the theoretical expectation that there will be variation in true expression even in control data[22] (Table 4). However, with the exception of the GemCode data, we found essentially no spike-in genes departed from a Gamma expression model. Our results broadly match those of ref.[17], that a NB observation model adequately captures the variation in the observed counts. However, we interpret these results differently: the good fit of an NB model to observed control data implies that the observation model is NB, not that the measurement model is NB. Rather, the NB observation model is consistent with a Poisson measurement model and a Gamma expression model.

In biological data sets, we found most genes (>94%) did not depart significantly from a Gamma expression model ($p \geq 0.05$ after Bonferroni correction; Figure 1B). Surprisingly, this was true even for *in silico* mixtures of distinct cell types, and for heterogeneous tissue

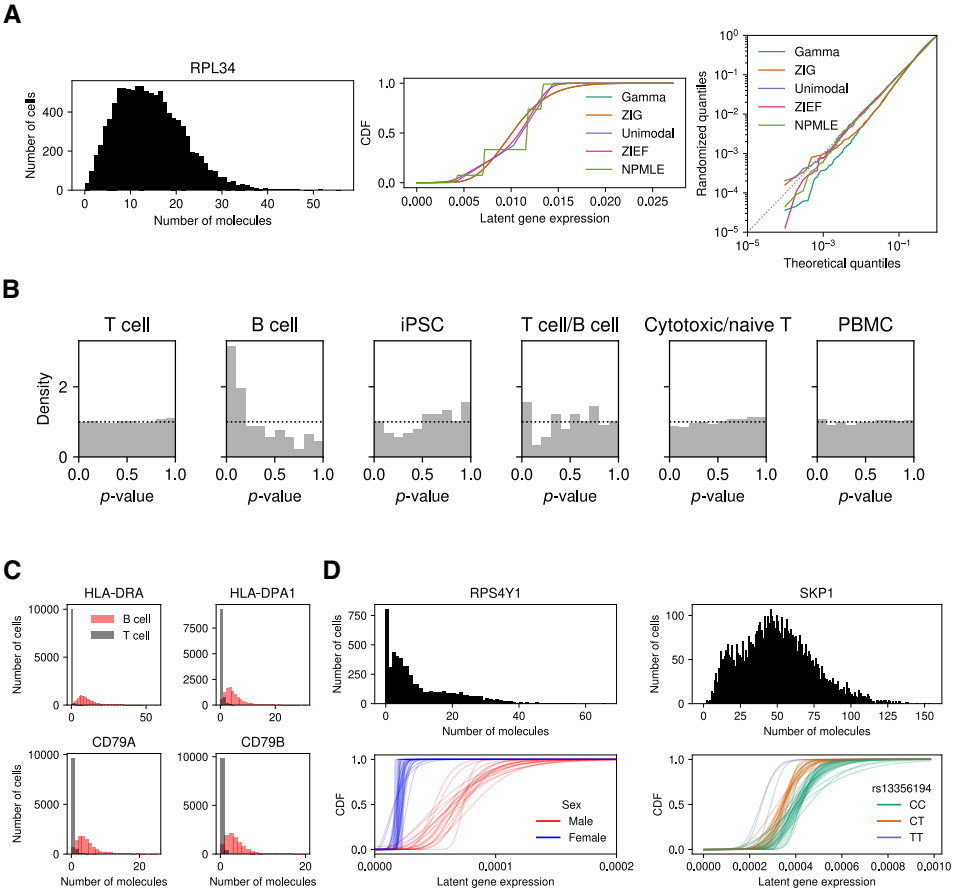

FIGURE 1. **Goodness of fit of single gene models of scRNA-seq.** (A) Example fits of different expression models to observed data at a single gene, and QQ plot of randomized quantiles of the estimated distributions against uniform quantiles. The data depart from the fitted distribution if the randomized quantiles are significantly different from uniform quantiles. ZIG: mixture of point mass on zero and Gamma; ZIEF: mixture of point mass on zero and exponential family; NPMLE: fully nonparametric distribution. (B) Histogram of p-values for departure from the fitted distribution for each gene in each data set, assuming a Gamma expression model. (C) Top genes in an *in silico* mixture of B cells and T cells departing from the fitted Gamma expression model. (D) Example genes in iPSCs showing complex marginal distributions of gene expression due to donor-specific effects.

samples. Further, we observed an enrichment of small goodness-of-fit p-values for gene expression in sorted B cells, which was not observed in any other data set.

There are multiple ways that a gene's true expression levels could deviate from a Gamma distribution: for example, the expression could be a mixture of zero and non-zero components, for which a point-Gamma model might be necessary; or the expression could follow

Dataset	Point mass	Gamma	Point-Gamma	Unimodal
Chromium (rep. 1)	9	3	3	0
Chromium (rep. 2)	14	0	0	0
DropSeq	65	0	0	0
GemCode	28	20	21	0
InDrops	2	0	0	0

TABLE 4. **Goodness of fit of different single-gene expression models in control scRNA-seq data.** Number of genes which significantly depart from the estimated expression model ($p < 0.05$ after Bonferroni correction).

a more complex multi-modal distribution; or, more simply, the expression variation could simply be larger than predicted by a Gamma. To assess which of these deviations might be more common we focused on those genes that departed significantly from a Gamma expression model, and asked whether a point-Gamma expression model adequately fit the data.

First, we analyzed control data and found that no genes were better fit by a point-Gamma expression model (Table 4). This result provides strong empirical support for the claim that zero-inflated measurement models are not needed for scRNA-seq data, because it shows that adding a point mass on zero (to either measurement or expression model) is generally not helpful in improving the fit to the data.

We then analyzed biological data sets, and found that in most cases only a minority (6–22%) of the genes that departed from a Gamma expression model were adequately fit by the point-Gamma model. These results shows that, in most cases, departures from a Gamma expression model in these data are not simply explained by an excess of zeroes. The only exception was the T cell/B cell mixture, where 53% of genes that departed from a Gamma expression model were adequately fit by a point-Gamma expression model. In this case, there are genes known not to be expressed in half of the cells, and those genes clearly require a point mass on zero in the expression model (Figure 1C). This result illustrates that our goodness-of-fit procedure is capable of finding such patterns when they exist.

These results suggest that bimodal and multi-modal gene expression among cells could be rarer than previously suggested[42–44]. To directly assess evidence for multi-modal gene expression, we tested for the goodness of fit of a unimodal (non-parametric) expression model, which can be fit using *generalized adaptive shrinkage*[25,29] (Appendix B).

In control experiments, for the minority of genes that departed from a Gamma expression model, we found no genes significantly depart from a unimodal expression model (Table 4). This result suggests a possible reason previous studies found a Poisson distribution inadequate to capture measurement error in some control data sets[17,22]: the expression models considered may have simply not been flexible enough.

In the non-mixture biological data sets, we found that 0–5% of genes that departed from a Gamma expression model also depart from a unimodal expression model (with the exception of the B cells data set). In B cells, we found 78% of genes departed, and in the mixture data sets, we found 50–68% of genes departed. However, on inspection of the distribution of observed molecule counts, many of these cases were clearly unimodal, with mode near zero, suggesting that they reflect a failure of the generalized adaptive shrinkage method. One potential explanation is that accurately estimating the density near zero of a unimodal non-parametric distribution constrained to non-negative reals is difficult[29].

The fact that almost all genes show unimodal expression does not imply that there is not considerable heterogeneity in expression among cells, or indeed substructure among samples. To illustrate this point, we focus on two examples from the iPSC data that exhibit complex marginal distributions of molecule counts explained by donor-specific distributions of true gene expression (Figure 1D). The study that generated these data fit a point-Gamma expression model at each gene within each individual, and found that only 0.01% of individual-gene combinations significantly departed from the estimated distribution[40]. *RPS4Y1* is a Y-chromosome linked gene, which exhibits clearly distinct distributions of gene expression for male and female donors. (The reason cells from female donors are estimated to have non-zero expression of *RPS4Y1* is likely because the relevant part of the coding sequence is identical between *RPS4Y1* and its homolog *RPS4X* and some reads erroneously mapped.) However, even within male donors, there is variation in the mode due to differing genetic background between individuals. Further highlighting the impact of differing genetic background on expression variation, *SKP1* is a gene whose expression is linked to the genotype of a nearby SNP, and analyzing cells from each donor separately reveals different modes among donors both between and within genotype classes.

Overall, our results suggest that a Gamma expression model may be adequate for the majority of genes in many data sets, and that a unimodal expression model captures most of the exceptions. One important implication of our results is that simpler, analytically tractable models for scRNA-seq data could be sufficient for many analysis tasks that are usually performed one gene at a time, e.g., differential expression analysis.

4.2. **Multi-gene models.** In multi-gene models, a common goal is to estimate the underlying low rank structure of the expression matrix $\mathbf{\Lambda}$ (or, in models that include noise, the structured component of expression $\boldsymbol{\mu}$). For example, the structure of the expression matrix could provide biological insights into differences in the cell type/state of different samples. Assessing the effectiveness of different methods for providing such biological insights is important, but also difficult to do objectively. Therefore, as a proxy we instead attempted to assess multi-gene models in their ability to accurately estimate the underlying expression matrix, $\mathbf{\Lambda}$. Our goal here is to address questions such as whether non-linear models provide a better fit to scRNA-seq than linear models, and whether the use of a specific link function produces a consistently improved fit.

In real data sets, the true value of $\mathbf{\Lambda}$ is unknown. To side-step this problem, we used *binomial thinning*[45,46] of real data sets (independently proposed as *molecular cross-validation*[47]) to produce training and validation data with the same true expression matrix $\mathbf{\Lambda}$. The key idea of this approach is that a Binomial sample of a Poisson-distributed count is also (marginally) Poisson-distributed, meaning it can produce such training/validation data without requiring knowledge of the true $\mathbf{\Lambda}$. We then assessed multi-gene methods by fitting them to the training data, and measuring their predictive accuracy (Poisson log likelihood) on the validation data. The assumption here is that models that more accurately capture the structure in the data should have higher validation log likelihood.

We applied binomial thinning to several data sets (Table 3) and compared three different methods based on different noiseless multi-gene expression models: a linear model assuming identity link (NMF), a linear model assuming log link (GLM-PCA), and a non-linear model (Poisson variational autoencoder, PVAE; Appendix C). To facilitate comparisons, we implemented these methods within a single Python package *scmodes*, and applied them to sub-samples of 1,000 cells in each data set. We restricted our analysis to genes with non-zero observations in at least 25% of cells (this filter includes all genes in the iPSC data, so we sub-sampled to 500 random genes).

In these comparisons, NMF often performed worse on validation log likelihood than GLM-PCA and PVAE (Figure 2A). However, this difference was almost completely explained by behavior in fitting the entries equal to zero in the training data: NMF systematically estimated smaller λ_{ij} for these zero entries, leading to poor validation log-likelihood in cases where the validation set contained a non-zero count (Figure 2B). The Poisson log likelihood is particularly sensitive to small absolute differences (but large relative differences) near 0. Effectively, NMF has a greater tendency than the other methods to overfit the zero entries and underestimate the true expression values. While this behavior of NMF is statistically undesirable, it could likely be remedied by minor changes (e.g., regularization), and likely does not reflect a fundamental limitation of its assumptions (linear dimension reduction with identity link). Furthermore, it is unclear whether this behavior actually hinders the ability of NMF to capture the important biological structure in the data. Indeed, overall the three methods produce highly concordant estimates of expression $\mathbf{\Lambda}$ (Figure 2C).

In summary, these comparisons highlight a key idea: that multi-gene methods can be thought of as estimating the underlying expression matrix $\mathbf{\Lambda}$ (or in models with noise, $\boldsymbol{\mu}$) from observed counts. However, the validation log likelihood for estimating $\mathbf{\Lambda}$ may ultimately not be the most relevant measure of the utility of the model fit. Other metrics, such as clustering performance in the estimated low dimensional subspace[15,48,49] may be more helpful, although they require labeled data. Ultimately, it may be the case that practical issues (such as convergence behavior of specific algorithms and the computation time and memory required to fit the model), or subjective measures (such as ease of interpretation of the fitted model), may determine which methods are most useful in practice. Assessing expression models on these other metrics will be an important direction for future work.

5. Conclusion

Here, we described how models for observed scRNA-seq counts can be helpfully separated into two parts: measurement models that describe variation introduced by the measurement process, and expression models that describe variation in actual expression levels. We argued that a simple Poisson model suffices to describe the measurement model, and that many existing scRNA-seq observation models can be interpreted as combining the Poisson measurement model with different expression models. We explained how these two simple ideas help clarify confusion about the source and interpretation of zeros in scRNA-seq data, and give rigorous procedures to interrogate variation in gene expression among cells.

Although the framework outlined here encompasses a number of commonly used models, there are also some models and methods it does not include. In particular, a common (and often computationally advantageous) strategy is to first transform the count data, using a transformation such as $y_{ij} = \log(x_{ij}/x_{i+} + \epsilon)$, and then apply Gaussian methods such as principal components analysis[50], factor analysis[51], or latent variable models[52,53] to the transformed data $\mathbf{Y}$. Although such methods do not directly fit into the framework outlined here, which models the observed counts rather than transformations, some of the key ideas may nonetheless be useful when interpreting these methods. The transformed observations are still generated by the same basic processes (variation in gene expression among cells and measurement error) and it can be helpful to think about how the transformed data should relate to these processes. One potential starting point is Taylor series approximations[54,55], which yield the approximation $E[y_{ij}] \approx \log(\lambda_{ij} + \epsilon)$ and suggest that Gaussian low rank models of the form $E[\mathbf{Y}] = \mathbf{LF}'$ can be interpreted as assuming that $[\log(\lambda_{ij} + \epsilon)]$ is

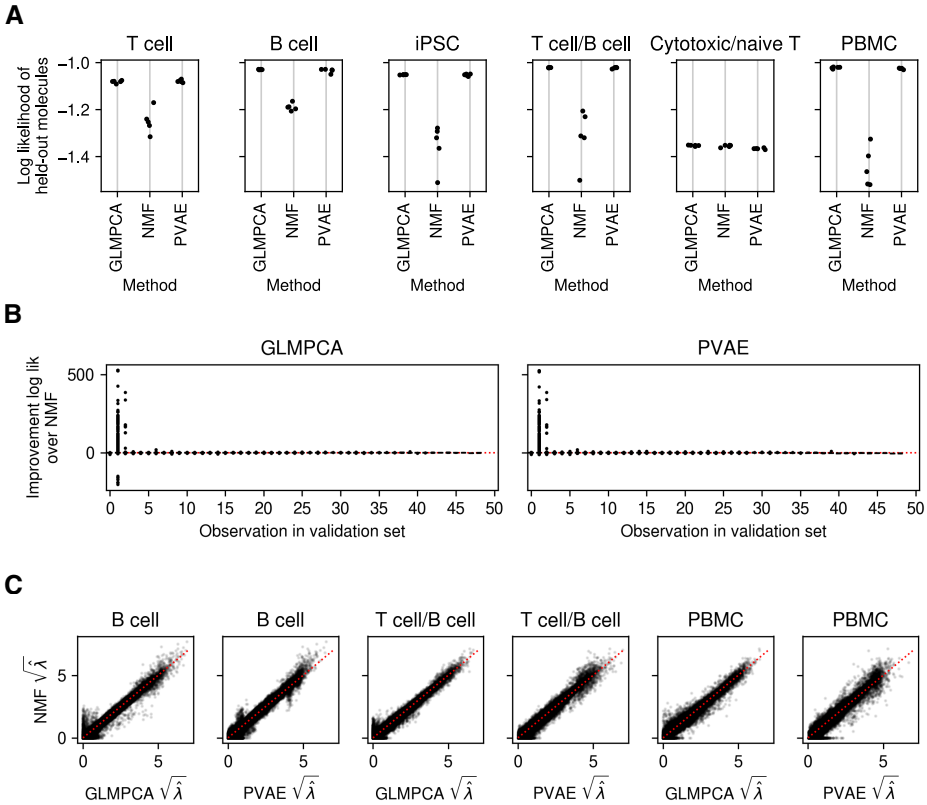

FIGURE 2. **Goodness of fit of multi-gene models of scRNA-seq.** (A) Per observation (sample-gene combination) log likelihood of held out molecules for 5 random training/validation splits, fixing the rank of each model to 8. (B) Distribution of log likelihood differences on one validation data set generated by holding out molecules from the B cells data set. (C) Comparison of estimated $\hat{\mathbf{\Lambda}}$ for three representative training data sets, fixing the rank of each model to 8.

low rank. A key problem moving forward will be to make such connections more rigorous, including assessments of the accuracy of such approximations, and how to best exploit the theoretical arguments behind the Poisson measurement model when applying Gaussian observation models to transformed data.

Our framework also highlights new directions for methods development, by revealing a space of expression models that has not been fully explored (in terms of e.g., low rank assumptions/transformations, noise models, or priors/constraints). Important problems to solve moving forward include: (1) efficiently learning complex gene expression distributions from heterogeneous samples, (2) making robust statistical comparisons between distributions for tasks such as differential mean expression or differential variance, (3) recovering interpretable loadings and factors from low rank expression models, and (4) efficiently fitting low rank expression models to massive, sparse data sets. Making progress on these problems will have important theoretical and practical relevance for the field.

URLs. Companion website https://aksarkar.github.io/singlecell-modes; analysis code https://www.github.com/aksarkar/singlecell-modes; *scmodes* Python package https://www.github.com/aksarkar/scmodes

Acknowledgements. We thank members of the Stephens and Gilad labs for helpful comments.

Appendix A. Poisson model for measurement error

A.1. **Derivation of basic Poisson measurement model.** Let m_{ij} denote the number of mRNA molecules present in cell i from gene j ($i = 1, \dots, n; j = 1, \dots, p$), and let $m_{i+} \triangleq \sum_j m_{ij}$. The simplest model for measurement error is to assume that the measured molecules in each cell are a simple random sample of all available molecules (sampled without replacement through the use of UMIs). Under this assumption, the observed molecule counts for each gene in sample i, $x_{i1}, \dots, x_{ip}$, follow a Multivariate Hypergeometric distribution. Unfortunately, the Multivariate Hypergeometric distribution is difficult to work with. However, if the number of molecules sampled $x_{i+} \triangleq \sum_j x_{ij}$ is small compared with the total number of molecules in the cell ($x_{i+} \ll m_{i+}$), then it is reasonable and convenient to approximate the Multivariate Hypergeometric distribution with the simpler Multinomial distribution (i.e., assuming sampling with replacement):

$$x_{i1}, \dots, x_{ip} \mid x_{i+}, \lambda_{i1}, \dots, \lambda_{ip} \sim \text{Multinomial}(x_{i+}; \lambda_{i1}, \dots, \lambda_{ip}), \tag{4}$$

where $\lambda_{ij} \triangleq m_{ij}/m_{i+}$ is the relative expression of gene j in sample i. Furthermore, it is often still more convenient to transform (4) into a Poisson model[3,56]:

$$x_{ij} \mid x_{i+}, \lambda_{ij} \sim \text{Poisson}(x_{i+}\lambda_{ij}), \qquad j = 1, \dots, p. \tag{5}$$

It is straightforward to show that the models (4) and (5) are equivalent, meaning they produce the same (log) likelihood for $\lambda_{i\cdot}$ up to a constant, which does not affect either maximum likelihood or Bayesian inference. Let l_{pois} and l_{mult} denote the log-likelihoods for $\lambda_{i\cdot}$ under (5) and (4), respectively. Then,

$$\begin{aligned}
l_{\text{pois}}(\lambda_{i\cdot}) &\triangleq \ln p(x_{i1}, \dots, x_{ip} \mid x_{i+}, \lambda_{i\cdot}) && (6)\\
&= \sum_j \left[x_{ij} \ln(x_{i+}\lambda_{ij}) - x_{i+}\lambda_{ij}\right] + \text{const} && (7)\\
&= \sum_j x_{ij} \ln \lambda_{ij} + \text{const} \qquad (\text{because } \textstyle\sum_j \lambda_{ij} = 1) && (8)\\
&= l_{\text{mult}}(\lambda_{i\cdot}) + \text{const}. && (9)
\end{aligned}$$

A.2. **Incorporating biased sampling.** The derivation above assumed random sampling of molecules, with all molecules equally likely to be sampled. However, the Poisson model can also accomodate biases in capture (library preparation) or sequencing that might make some molecules more likely to be sampled than others. Suppose the probability of successfully sampling a molecule from gene j is proportional to b_j, and suppose the measurement process generates each measured molecule by repeatedly drawing from the pool of molecules present until the first success. Then, for each molecule

$$\Pr(\text{measured molecule from gene } j \mid \text{measurement succeeded}) = \frac{b_j m_{ij}}{\sum_t b_t m_{it}}. \tag{10}$$

Without loss of generality, scale $b_1, \dots, b_p$ such that $\sum_j b_j m_{ij} = m_{i+}$. Then,

$$x_{i1}, \dots, x_{ip} \mid x_{i+}, \lambda_{i1}, \dots, \lambda_{ip}, b_1, \dots, b_p \sim \text{Multinomial}(x_{i+}; b_1\lambda_{i1}, \dots, b_p\lambda_{ip}), \tag{11}$$

implying

$$x_{ij} \mid x_{i+}, \lambda_{ij}, b_j \sim \text{Poisson}(x_{i+} b_j \lambda_{ij}), \qquad j = 1, \dots p. \tag{12}$$

A similar argument was previously made in modeling the impact of gene length on the measurement model for bulk RNA-seq read counts[3]. Here, we leave aside the question of estimating b_j from scRNA-seq UMI data, which has been previously discussed[22]. However,

we argue it is reasonable to assume that the biases b_j are generally small; otherwise, scRNA-seq measurements would be greatly devalued since they would reflect bias more than true gene expression.

A.3. **Violations of model assumptions.** The derivations above used a Multinomial (Poisson) model to approximate the Multivariate Hypergeometric distribution, which relies on the assumption $x_{i+} \ll m_{i+}$. We now briefly consider what could happen if this assumption is violated; i.e., if one actually captured and sequenced a substantial fraction of all available molecules.

Consider gene j, and partition the molecules into those which map to gene j and those which do not. Then, the Multivariate Hypergeometric distribution reduces to the Hypergeometric distribution, and the Multinomial distribution reduces to the Binomial distribution. Suppose one sequences a fraction p of the molecules, and consider the following candidate models:

$$x_{ij} \sim \text{Poisson}(m_{ij}p) \triangleq P_1 \tag{13}$$

$$x_{ij} \sim \text{Binomial}(m_{i+}p, m_{ij}/m_{i+}) \triangleq P_2 \tag{14}$$

$$x_{ij} \sim \text{Hypergeometric}(m_{i+}p, m_{i+}, m_{ij}) \triangleq P_3. \tag{15}$$

These models all have the same expectation, $E[x_{ij}] = m_{ij}p$, but unequal variances:

$$V_{P_1}[x_{ij}] = m_{ij}p \tag{16}$$

$$V_{P_2}[x_{ij}] = m_{ij}p\frac{m_{i+} - m_{ij}}{m_{i+}} \tag{17}$$

$$V_{P_3}[x_{ij}] = m_{ij}p\frac{m_{i+} - m_{ij}}{m_{i+}}\frac{m_{i+}}{m_{i+} - 1}(1 - p). \tag{18}$$

It is reasonable to assume that m_{i+} is large ($m_{i+} >> m_{ij}$ and $m_{i+} >> 1$), in which case both $(m_{i+} - m_{ij})/m_{i+}$ and $m_{i+}/(m_{i+} - 1)$ are close to one. Therefore, the difference in variance between the Hypergeometric and the Poisson model is primarily due to the factor $1-p$. As p becomes large the Poisson model has larger variance than the Hypergeometric distribution. Therefore, the Poisson model will, in principle, tend to overstate the amount of measurement error, and inferences based on Poisson models for measurement error will tend to understate expression variation. To what extent this will matter in practice as technologies develop to sample larger fractions of molecules remains to be seen.

Appendix B. Inference in single gene models

To analyze observed data $x_{1j}, \dots, x_{nj}$ at a single gene j, we combine the Poisson measurement model with an expression model g_j

$$x_{ij} \mid x_{i+}, \lambda_{ij} \sim \text{Poisson}(x_{i+}\lambda_{ij}) \tag{19}$$

$$\lambda_{ij} \sim g_j(\cdot) \in \mathcal{G}, \tag{20}$$

where $\mathcal{G}$ is some family of distributions, and estimate g_j from the data. This is achieved by maximizing the likelihood

$$L(g_j; x_{1j}, \dots, x_{nj}, x_{1+}, \dots, x_{n+}) \triangleq \prod_{i=1}^{n} \int_0^\infty \text{Poisson}(x_{ij}; x_{i+}\lambda_{ij})\, dg_j(\lambda_{ij}) \tag{21}$$

$$\hat{g}_j = \arg\max_{g_j \in \mathcal{G}} L(g_j; x_{1j}, \dots, x_{nj}, x_{1+}, \dots, x_{n+}). \tag{22}$$

Readers who are more familiar with likelihoods for parametric distributions should not be put off by our notation $L(g_j)$ here, which is simply the usual definition of likelihood, but

with the "parameter" being a distribution. If $\mathcal{G}$ is a parametric family, parameterized by $\theta \in \Theta$ say, then maximizing $L(g_j)$ over $g_j \in \mathcal{G}$ is simply the same as maximizing $L(\theta)$ over $\theta \in \Theta$. We use the more general notation to allow that $\mathcal{G}$ could be a non-parametric family.

Different assumptions on $\mathcal{G}$ trade off ease of implementation and computational cost for flexibility and modeling power. If one makes strong parametric assumptions about $\mathcal{G}$, then the estimation procedures are easy. For example, if one assumes g_j is a Gamma distribution, then the integrals in (21) are analytic, and yield Negative Binomial distributions. (We highlight a subtle point: since the distribution of measurement error for observation x_{ij} depends on x_{i+}, under this model each observation comes from its own specific distribution. The individual distributions share only some parameters, depending on the chosen family $\mathcal{G}$.) Similarly, if one assumes g_j is a mixture of a point mass on zero and a Gamma distribution, then the integrals yield Zero-inflated Negative Binomial distributions. In the simplest cases, these observation models can be written as generalized linear models, and can be readily fit using standard packages[24]. Additionally, many published methods for both bulk and single cell RNA-seq data have adopted the NB observation model, although they do not necessarily directly perform downstream inferences on the corresponding Gamma expression model.

If one interprets g_j as a prior and $p(x_{ij} \mid x_{i+}, \lambda_{ij})$ as a likelihood, then the inference procedure is empirical Bayes. This insight suggests an approach based on *generalized adaptive shrinkage*[25,29]. The key idea of generalized adaptive shrinkage is to assume g_j is a unimodal (non-parametric) distribution on non-negative reals. In practice, g_j is represented as a large, but finite, mixture of uniform distributions

$$g_j(\cdot) = \sum_{k=1}^{K_j} w_{jk} \,\mathrm{Uniform}(\cdot; \lambda_{0j}, a_{jk}), \tag{23}$$

where the number of endpoints K_j and their locations a_{jk} are fixed based on the range of the data. (Here $\mathrm{Uniform}(\cdot; \lambda_{0j}, a_{jk})$ denotes the density of the uniform distribution on $[a_{jk}, \lambda_{0j}]$ if $a_{jk} < \lambda_{0j}$, and on $[\lambda_{0j}, a_{jk}]$ if $a_{jk} > \lambda_{0j}$.) Assuming the mode λ_{0j} is known, the likelihood (21) is a function only of the mixture weights $\mathbf{w}_j$, resulting in a convex optimization problem. To estimate λ_{0j}, we use Brent's method to optimize (21) with respect to λ_{0j}, at each candidate point optimizing over mixture weights $\mathbf{w}_j$. This approach is implemented in the R package *ashr*.

Another flexible approach, which assumes g_j is a mixture of a point mass on zero and an exponential family, was recently proposed[22,57]. In this model,

$$g_j(\cdot) = \pi_j \delta_0(\cdot) + (1 - \pi_j) \exp\left(\mathbf{Q}(\cdot)'\boldsymbol{\alpha} - \phi(\boldsymbol{\alpha})\right) \tag{24}$$

where $\mathbf{Q}(\cdot)$ is the degree 5 natural cubic spline basis, $\boldsymbol{\alpha}$ is the 5×1 natural parameter, and $\phi(\cdot)$ is the normalizer. In practice, the basis $\mathbf{Q}$ and the representation of the distribution g_j are both discretized, and a regularization penalty is added to the likelihood (21). This approach is flexible enough to represent many unimodal distributions and some multi-modal distributions due to its use of (smooth) natural splines, and is implemented in the R package *DESCEND*.

The most general approach is to assume that g_j is simply some (non-parametric) distribution on non-negative reals[30,58]. Any such distribution g_j can be approximated by a mixture of uniforms on a dense grid

$$g_j(\cdot) = \sum_{k=1}^{K_j} w_{jk} \mathrm{Uniform}(\cdot; (k-1)a_j, ka_j), \tag{25}$$

where the number of components K_j depends on the range of the data and the grid step size a_j is fixed. This model reduces to the same convex optimization problem as (23), and is implemented in the R package *ashr*.

To test whether a particular distributional assumption is sufficient to model expression variation at each gene, we applied a goodness-of-fit test based on a simple fact: if the data were truly distributed according to some cumulative distribution function F, then the values of F evaluated at the data will be uniformly distributed between 0 and 1. To use this idea, we estimated $g_j \in \mathcal{G}$ under the assumed $\mathcal{G}$, which yields an estimate of the marginal distribution for each observation $\hat{F}_{ij}$ according to (21). We then applied the Kolmogorov-Smirnov test to test whether the values $\hat{F}_{1j}(x_{1j}), \ldots, \hat{F}_{nj}(x_{nj})$ significantly depart from the uniform distribution. (This procedure has to be randomized to account for the fact that for discrete count data, the CDF F is discontinuous[59].)

Appendix C. Inference in multi-gene models

To model variation in λ_{ij} across both cells and genes simultaneously, one approach that is both common and powerful is to assume that the matrix $\mathbf{\Lambda}$ (or some transformation of it) is low rank. This *structural assumption* can be written

$$\lambda_{ij} = h^{-1}((\mathbf{LF}')_{ij}), \tag{26}$$

where h is a *link function* (for example, h could be the identity function or log function), $\mathbf{L}$ is an $n \times k$ matrix of *loadings*, and $\mathbf{F}$ is a $p \times k$ matrix of *factors*. One could in principle incorporate a link function in a single gene model $h(\lambda_{ij}) \sim g_j(\cdot) \in \mathcal{G}$; however, doing so would be exactly equivalent to transforming the distribution g_j, and so would not provide any additional generality. In contrast, assuming h is identity in (26) is fundamentally different from assuming $h = \log$: if the matrix $[\lambda_{ij}]$ is rank k, it is not necessarily true that the matrix $[\log \lambda_{ij}]$ is also rank k (or vice versa).

If one only makes the structural assumption (26), then one assumes that the only variation in the data beyond variation in the loadings $l_{1\cdot}, \ldots, l_{n\cdot}$ is measurement error. This assumption is clearly restrictive, and it is natural to instead assume that expression variation is partitioned into structured and unstructured (stochastic) components, corresponding to a structural assumption and a distributional assumption, respectively:

$$\begin{aligned} \lambda_{ij} &= \mu_{ij} u_{ij} \\ \mu_{ij} &= h^{-1}((\mathbf{LF}')_{ij}) \\ u_{ij} &\sim p_u(u_{ij}) \end{aligned} \tag{27}$$

where u_{ij} is a multiplicative random effect. Marginalizing over u_{ij} yields $\lambda_{ij} \sim g_{ij}(\cdot) \in \mathcal{G}$, connecting single-gene and multi-gene expression models in our framework. This model can be further augmented by assuming that u_{ij} has structure (for example, gene-specific and/or sample-specific variance) which is not of interest. Fitting this class of models to data involves the same two steps as fitting single gene models: (1) estimating g, and (2) estimating λ_{ij}.

C.1. **Zero-inflated Negative Binomial models.** As concrete examples, we show how to write ZINB-WaVE[39], DCA[23], and scVI[15] as instances of our framework. All three methods arise from a point-Gamma assumption

$$u_{ij} \sim \pi_{ij}\delta_0(\cdot) + (1 - \pi_{ij})\,\mathrm{Gamma}(\phi_{ij}^{-1}, \phi_{ij}^{-1}), \tag{28}$$

which implies $\mathcal{G}$ is also the family of point-Gamma distributions (because this family is closed under multiplication). ZINB-WaVE makes the following structural assumptions:

$$\ln \mu_{ij} = \mathbf{L}\mathbf{F}'_{\mu} \tag{29}$$

$$\phi_{ij} = \phi_j \tag{30}$$

$$\operatorname{logit} \pi_{ij} = \mathbf{L}\mathbf{F}'_{\pi}, \tag{31}$$

where for brevity we have omitted fixed effects for sample- and gene-specific covariates. In this model, the matrices $\mathbf{F}_{\mu}$ and $\mathbf{F}_{\pi}$ are linear mappings from k-dimensional latent variables $l_{i\cdot}$ to p-dimensional vectors $\mu_{i\cdot}$ and $\pi_{i\cdot}$. ZINB-WaVE estimates $\mathbf{L}, \mathbf{F}_{\mu}, \mathbf{F}_{\pi}, \phi_1, \ldots, \phi_p$ by maximizing a penalized log likelihood.

scVI and DCA each use a neural network to represent a non-linear mapping from $x_{i\cdot}$ to $l_{i\cdot}$, and from $l_{i\cdot}$ to e.g., $\mu_{i\cdot}$. DCA is a classical autoencoder in which $l_{i\cdot}$ is the output of the *bottleneck layer*, and the low rank assumptions are:

$$\mu_{i\cdot}(l_{i\cdot}) = \exp(\mathbf{W}_{\mu} \operatorname{ReLU}(\mathbf{W}_{\mathbf{L}} l'_{i\cdot})) \tag{32}$$

$$\phi_{i\cdot}(l_{i\cdot}) = \exp(\mathbf{W}_{\phi} \operatorname{ReLU}(\mathbf{W}_{\mathbf{L}} l'_{i\cdot})) \tag{33}$$

$$\pi_{i\cdot}(l_{i\cdot}) = \operatorname{sigmoid}(\mathbf{W}_{\pi} \operatorname{ReLU}(\mathbf{W}_{\mathbf{L}} l'_{i\cdot})) \tag{34}$$

where $\operatorname{sigmoid}(x) = 1/(1 + \exp(-x))$; $\operatorname{ReLU}(x) = \max(0, x)$; functions are applied element-wise; $\mathbf{W}_{\mathbf{L}}$ is an $m \times k$ matrix, where $k < m < p$; and $\mathbf{W}_{\mu}, \mathbf{W}_{\phi}, \mathbf{W}_{\pi}$ are $p \times m$ matrices. DCA minimizes the ZINB negative log likelihood.

scVI is a variational autoencoder[60], in which an *encoder network* $q(l_{i\cdot} \mid x_{i\cdot})$ represents the mapping from $x_{i\cdot}$ to $l_{i\cdot}$ as an approximate posterior distribution

$$q(l_{i\cdot} \mid x_{i\cdot}) = \operatorname{Normal}(m(x_{i\cdot}), \operatorname{diag}(s^2(x_{i\cdot}))) \tag{35}$$

where $m(\cdot)$ and $s^2(\cdot)$ are k-dimensional neural network outputs representing the mean and diagonal of the covariance matrix, and a *decoder network* represents the mapping from $l_{i\cdot}$ to the distribution of $\lambda_{i\cdot}$ as outputs $\mu_{i\cdot}(l_{i\cdot})$ and $\pi_{i\cdot}(l_{i\cdot})$. In the simplest case, scVI also makes an assumption that $\phi_{ij} = \phi_j$. (We highlight a subtle point: scVI treats the effect of size factors in the measurement model as a random effect whose variance is estimated from the data, where in our discussion we have assumed it is a fixed effect of known magnitude.) Rather than maximizing a likelihood, scVI maximizes a lower bound to the model evidence.

Importantly, this formulation implies that in all of these models, the mean gene expression is $(1 - \pi_{ij})\mu_{ij}$, and the distribution of λ_{ij} given the data is

$$\lambda_{ij} \mid x_{ij}, x_{i+}, \hat{g}_{ij} \sim \hat{\pi}_{ij}\delta_0(\cdot) + (1 - \hat{\pi}_{ij}) \operatorname{Gamma}(x_{ij} + \hat{\phi}_{ij}^{-1}, x_{i+} + \hat{\phi}_{ij}^{-1}\hat{\mu}_{ij}^{-1}), \tag{36}$$

which is not how these methods have been derived or interpreted.

C.2. **Poisson variational autoencoder.** In Section 4.2, we introduced PVAE, which is written

$$x_{ij} \mid l_{i\cdot} \sim \operatorname{Poisson}((\lambda(l_{i\cdot}))_j) \tag{37}$$

where $\lambda(\cdot)$ is a decoder neural network mapping k-vector $l_{i\cdot}$ to p-vector $\lambda_{i\cdot}$. In this approach, we assume no noise. The encoder network is given by (35). We used two hidden layers of dimension 128 in the encoder and decoder networks, ReLU activations, and batch normalization at each hidden layer. We fit the model using batch gradient descent accelerated by RMSprop[61]. We held out 10% of the training data as a test set to determine early

stopping criteria for the optimization. After fitting the model, we estimated

$$\hat{\lambda}_{ij} = \frac{1}{S}\sum_{s=1}^{S}(\lambda(l_{i\cdot}^{(s)}))_j, \qquad l_{i\cdot}^{(s)} \sim q(l_{i\cdot} \mid x_{i\cdot}). \tag{38}$$

Appendix D. Identifiability of measurement and expression models

Multiple combinations of measurement model and expression model can give rise to the same observation model. Despite the fact that the resulting observation models are the same, and therefore cannot be distinguished from observed data alone, different choices for expression models will naturally give different inferences about true gene expression. We illustrate this fact using the ZINB observation model as an example. The ZINB observation model for counts $x_{1j}, \dots, x_{nj}$ observed at a single gene j is a mixture of two distributions, a point mass at 0 and a Negative Binomial distribution. In our framework, it arises by assuming a point-Gamma expression model (a mixture of a point mass at 0, denoted δ_0, and a Gamma distribution)

$$\begin{aligned} x_{ij} \mid \lambda_{ij}, x_{i+} &\sim \text{Poisson}(x_{i+}\lambda_{ij}), \qquad i = 1, \dots, n \\ \lambda_{ij} \mid \mu_j, \phi_j, \pi_j &\sim \pi_j \delta_0(\cdot) + (1-\pi_j)\,\text{Gamma}(\phi_j^{-1}, \mu_j^{-1}\phi_j^{-1}). \end{aligned} \tag{39}$$

It is straightforward to show that marginalizing over λ_{ij} yields a ZINB distribution on $x_{1j}, \dots, x_{nj}$. In this parametrization, the Gamma component of the point-Gamma distribution has mean μ_j and variance $\mu_j^2\phi_j$, and the resulting Negative Binomial component of the ZINB model has mean $x_{i+}\mu_j$, dispersion ϕ_j, and variance $x_{i+}\mu_j + (x_{i+}\mu_j)^2\phi_j$.

Deriving the ZINB model this way clarifies that fitting the ZINB observation model to data corresponds to estimating the underlying point-Gamma expression model, removing the effect of measurement error. This point-Gamma model contains all of the information about expression variation needed to answer biological questions of interest. For example, under the model (39), the mean gene expression is $(1-\pi_j)\mu_j$.

However, the ZINB observation model can be derived in a different way, using a different measurement model and different expression model, which leads to a different interpretation. Specifically, one could assume a *Zero-inflated Poisson* measurement model (a mixture of a point mass at 0 and a Poisson distribution), and a Gamma expression model

$$\begin{aligned} x_{ij} \mid \lambda_{ij}, x_{i+}, \pi_j &\sim \pi_j \delta_0(\cdot) + (1-\pi_j)\,\text{Poisson}(x_{i+}\lambda_{ij}) \\ \lambda_{ij} \mid \mu_j, \phi_j &\sim \text{Gamma}(\phi_j^{-1}, \mu_j^{-1}\phi_j^{-1}). \end{aligned} \tag{40}$$

It is straightforward to show that marginalizing over λ_{ij} in (40) also leads to a ZINB distribution; in fact, the same ZINB distribution as (39).

Because these two different sets of assumptions, (39) and (40), produce exactly the same observation model, one cannot distinguish between them based on which better fits the observed data better. However, distinguishing between them matters in practice because they lead to different inferences about the actual expression values from the observed data. For example, whereas under (39) the mean of λ_j is $(1-\pi_j)\mu_j$, under (40) it is μ_j. More generally, these two different interpretations could lead to different inferences about e.g., differential expression or clustering[62]. We argue that both theory and empirical evidence support the use of the Poisson measurement model, and not the ZIP measurement model, and that therefore analyses using the ZINB observation model should be derived and interpreted using (39) rather than (40).

References

1. Fuller, W. A. *Measurement Error Models* (John Wiley & Sons, Inc., USA, 1986).
2. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18,** 1509–1517 (2008).
3. Pachter, L. Models for transcript quantification from RNA-Seq. *arXiv e-prints* (2011).
4. Wang, W. & Stephens, M. Empirical Bayes Matrix Factorization. *arXiv e-prints* (2018).
5. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11,** 740 (2014).
6. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16.** Review Article, 133 (2015).
7. Zhu, L., Lei, J., Devlin, B. & Roeder, K. A unified statistical framework for single cell and bulk RNA sequencing data. *Ann Appl Stat* **12,** 609–632 (2018).
8. 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).
9. Genotyping Errors with the Polymerase Chain Reaction. *N Engl J Med* **322,** 61–61 (1990).
10. Whale, A. S., Cowen, S., Foy, C. A. & Huggett, J. F. Methods for Applying Accurate Digital PCR Analysis on Low Copy DNA Samples. *PLoS ONE* **8,** 1–10 (2013).
11. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* (2017).
12. Van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174,** 716–729.e27 (2018).
13. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat Commun* **9,** 997 (2018).
14. Chen, M. & Zhou, X. VIPER: variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol* **19,** 196 (2018).
15. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15,** 1053–1058 (2018).
16. Talwar, D., Mongia, A., Sengupta, D. & Majumdar, A. AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Sci Rep* **8,** 16329 (2018).
17. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat Biotech* (2020).
18. Huang, M. *et al.* SAVER: gene expression recovery for single-cell RNA sequencing. *Nat Methods* **15,** 539–542 (2018).
19. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11,** 163 (2013).
20. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21,** 1160–1167 (2011).
21. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol* **20,** 295 (2019).
22. Wang, J. *et al.* Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc Natl Acad Sci USA* (2018).
23. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun* **10,** 390 (2019).
24. Hilbe, J. M. *Modeling Count Data* (Cambridge University Press, 2014).

25. Lu, M. *Generalized Adaptive Shrinkage Methods and Applications in Genomics Studies* PhD thesis (University of Chicago, 2018).
26. 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 (2009).
27. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15,** 550 (2014).
28. Vallejos, C. A., Marioni, J. C. & Richardson, S. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. *PLoS Comp Biol* **11,** 1–18 (2015).
29. Stephens, M. False discovery rates: a new deal. *Biostatistics* **18,** 275–294 (2017).
30. Kiefer, J. & Wolfowitz, J. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *Ann Math Statist* **27,** 887–906 (1956).
31. Zhu, X., Ching, T., Pan, X., Weissman, S. & Garmire, L. Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ,* 5:e2888 (2017).
32. Matsumoto, H. *et al.* An NMF-based approach to discover overlooked differentially expressed gene regions from single-cell RNA-seq data. *NAR Genomics and Bioinformatics* **2.** lqz020 (2019).
33. Lee, D. D. & Seung, H. S. *Algorithms for Non-negative Matrix Factorization* in *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA* (eds Leen, T. K., Dietterich, T. G. & Tresp, V.) (MIT Press, 2000), 556–562.
34. Cemgil, A. T. Bayesian Inference for Nonnegative Matrix Factorisation Models. *Comput Intell Neurosci* (2009).
35. Gouvert, O., Oberlin, T. & Févotte, C. Negative Binomial Matrix Factorization for Recommender Systems. *arXiv e-prints* (2018).
36. Raj, A. & van Oudenaarden, A. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* **135,** 216–226 (2008).
37. Levitin, H. M. *et al.* De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. *Mol Syst Biol* **15** (2019).
38. Sun, S., Chen, Y., Liu, Y. & Shang, X. A fast and efficient count-based matrix factorization method for detecting cell types from single-cell RNAseq data. *BMC Syst Biol* **13,** 28 (2019).
39. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* **9,** 284 (2018).
40. Sarkar, A. K. *et al.* Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genetics* **15,** 1–16 (2019).
41. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8,** 14049 (2017).
42. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498,** 236 (2013).
43. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510,** 363 (2014).
44. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol* **17,** 63 (2016).

45. Gerard, D. & Stephens, M. Unifying and Generalizing Methods for Removing Unwanted Variation Based on Negative Controls. *arXiv e-prints* (2017).
46. Gerard, D. Data-based RNA-seq Simulations by Binomial Thinning. *bioRxiv* (2019).
47. Batson, J., Royer, L. & Webber, J. Molecular Cross-Validation for Single-Cell RNA-seq. *bioRxiv* (2019).
48. Hu, Q. & Greene, C. S. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. *Pac Symp Biocomput* **24,** 362–373 (2019).
49. Sun, S., Zhu, J., Ma, Y. & Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol* **20,** 269 (2019).
50. Tipping, M. E. & Bishop, C. M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61,** 611–622.
51. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol* **16,** 241 (2015).
52. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotech* **33,** 155–160 (2015).
53. Verma, A. & Engelhardt, B. E. A robust nonlinear low-dimensional manifold for single cell RNA-seq data. *bioRxiv* (2018).
54. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15,** R29 (2014).
55. Lun, A. Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data. *bioRxiv* (2018).
56. Baker, S. G. The Multinomial-Poisson Transformation. *J Royal Stat Soc* **43,** 495–504 (1994).
57. Efron, B. Empirical Bayes deconvolution estimates. *Biometrika* **103,** 1–20 (2016).
58. Koenker, R. & Mizera, I. Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules. *J Am Stat Assoc* **109,** 674–685 (2014).
59. Dunn, P. K. & Smyth, G. K. Randomized Quantile Residuals. *J Comput Graph Stat* **5,** 236–244 (1996).
60. Kingma, D. P. & Welling, M. *Auto-Encoding Variational Bayes* in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* (eds Bengio, Y. & LeCun, Y.) (2014).
61. Tieleman, T. & Hinton, G. *Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude* COURSERA: Neural Networks for Machine Learning. 2012.
62. Silverman, J. D., Roche, K., Mukherjee, S. & David, L. A. Naught all zeros in sequence count data are the same. *bioRxiv* (2018).