Genome Biology

REVIEW

Open Access

Statistics or biology: the zero-inflation controversy about scRNA-seq data

Ruochen Jiang[1], Tianyi Sun[1], Dongyuan Song[2] and Jingyi Jessica Li[1,3,4,5]*

*Correspondence: jli@stat.ucla.edu
[1] Department of Statistics, University of California, Los Angeles 90095-1554, CA, USA
[3] Department of Human Genetics, University of California, Los Angeles 90095-7088, CA, USA
Full list of author information is available at the end of the article

Abstract

Researchers view vast zeros in single-cell RNA-seq data differently: some regard zeros as biological signals representing no or low gene expression, while others regard zeros as missing data to be corrected. To help address the controversy, here we discuss the sources of biological and non-biological zeros; introduce five mechanisms of adding non-biological zeros in computational benchmarking; evaluate the impacts of non-biological zeros on data analysis; benchmark three input data types: observed counts, imputed counts, and binarized counts; discuss the open questions regarding non-biological zeros; and advocate the importance of transparent analysis.

Introduction

The rapid development of single-cell technologies has brought unprecedented opportunities to quantifying transcriptome heterogeneity among individual cells and transcriptome dynamics along cell developmental trajectories [1–4]. Many single-cell RNA sequencing (scRNA-seq) protocols have been developed. Two major types of protocols are (1) tag-based, unique molecular identifier (UMI)-based protocols such as Drop-seq [5] and 10x Genomics Chromium [6, 7] and (2) full-length, non-UMI-based protocols such as Smart-seq2 [8] and Fluidigm C1 [9]. Different protocols exhibit disparate accuracy and noise levels for quantifying gene expression in single cells, posing many computational and analytical challenges for researchers to extract biological knowledge from scRNA-seq data [10, 11]. Facing these challenges, computational researchers have developed hundreds of computational and statistical methods for various scRNA-seq data analytical tasks, including the selection of informative marker genes [12–16], the identification of cell types and states [14, 17–23], the reconstruction of cell developmental trajectories [24–29], and the identification of cell-type-specific genes [13, 28, 30–38].

A universal analytical challenge for scRNA-seq data generated by any protocol is the vastly high proportion of genes with zero expression measurements in each cell. This data sparsity issue is apparent when scRNA-seq data are compared with bulk RNA-seq data [36, 39, 40], which contain aggregated gene expression measurements from many cells.

While the proportion of zeros in bulk RNA-seq data is usually 10–40% [41, 42], that proportion can be as high as 90% in scRNA-seq data [43]. Such excess zeros would bias the estimation of gene expression correlations [44] and hinder the capture of gene expression dynamics [45] from scRNA-seq data. In early scRNA-seq data analyses, the high data sparsity provoked the use of zero-inflated models [36, 38, 46] and the development of imputation methods for reducing zeros [20, 44, 45, 47–63]. More recently, however, there were voices against the use of zero-inflated models for scRNA-seq data generated by UMI protocols [64, 65]. Besides, there was a proposal for treating zeros as useful information that researchers should embrace [66]. These mixed statements raised a fundamental question to the scRNA-seq field: should we use or remove zeros in scRNA-seq data analysis?

In this article, we provide some perspectives on this puzzling question by discussing the sources of zeros in scRNA-seq data, the impacts of zeros on various data analyses, the existing approaches for handling zeros, and the pros and cons of these approaches. Specifically, first, we define biological and non-biological zeros arising from scRNA-seq data generation, and we clarify several ambiguous terms about zeros in the scRNA-seq literature. Second, we use scRNA-seq data generated by Drop-seq, 10x Genomics, and Smart-seq2 to demonstrate the relation between zero patterns and protocols. Further, we use simulation studies to evaluate the effects of zeros and zero-generation mechanisms on cell clustering and differentially expressed (DE) gene identification. Third, we summarize three commonly used approaches for handling zeros—direct statistical modeling, imputation, and binarization—and discuss their respective pros and cons. Fourth, we benchmark the performance of the three input data types in three downstream analyses: cell clustering, cell dimension reduction, and DE gene identification. Last, we provide practical advice and outline future directions for bioinformatics tool developers and users.

Table 1 summarizes the key concepts used in this paper, including their definitions and categories (biology, technology, and modeling). Table 2 clarifies three zero-related terms: dropouts, excess zeros, and zero inflation; the first two have been ambiguously used in the literature.

Sources of zeros in scRNA-seq data

Zero measurements in scRNA-seq data have two sources: biological and non-biological (Fig. 1). While biological zeros carry meaningful information about cell states, non-biological zeros represent missing values artificially introduced during the generation of scRNA-seq data. In our paper, non-biological zeros include technical zeros, which occur during the preparation of biological samples for sequencing, and sampling zeros, which arise due to limited sequencing depths. Our classification of zeros in sequencing data into biological, technical, and sampling zeros is aligned with the classification in Silverman et al. [67] except a slight difference (we refer to the zeros due to inefficient amplification, e.g., PCR, as sampling zeros, while Silverman et al. called them technical zeros). The non-biological zeros have typically been viewed as impediments to the full and accurate interpretation of cell states and the differences between them. Figure 1a provides an overview of a scRNA-seq experiment, and it highlights the biological factors and technical procedures that may lead to zeros in scRNA-seq data. Figure 1b summarizes how biological factors result in biological zeros and how technical procedures cause non-biological zeros, including technical zeros and sampling zeros. It is worth noting that biological

Table 1 A summary of the key concepts used in this paper, including their definitions and nature

Key concepts	Definition	Nature
RNA polymerase	An enzyme that transcribes a DNA sequence into an RNA sequence	Biology
mRNA degredation	The process of an mRNA sequence being destroyed	Biology
Biological zero	Absence of mRNA of a gene in a cell	Biology
GC-rich	Majority of the bases in a sequence are either cytosine (C) or guanine (G)	Biology
Reverse transcription	Enzyme-mediated synthesis of a DNA molecule from an RNA template; a step to enable DNA sequencing	Sequencing technology
cDNA	Complementary DNA (synthesized from reverse transcription)	Sequencing technology
PCR	Polymerase chain reaction; a step to amplify cDNA copy number	Sequencing technology
IVT	In vitro transcription amplification; a step to amplify cDNA copy number	Sequencing technology
Sequence read	A short sequence read out by sequencing machine	Sequencing technology
UMI	Unique molecular identifier, which is used to correct amplification bias	Sequencing technology
Non-biological zero	Absence of reads or UMIs of a gene in a cell in scRNA-seq data when the gene in fact has mRNAs in the cell	Sequencing technology
Technical zero	Absence of reads or UMIs of a gene in a cell due to the library-preparation steps (e.g., cDNA synthesis) before cDNA amplification	Sequencing technology
Sampling zero	Absence of reads or UMIs of a gene in a cell due to inefficient amplification and/or limited sequencing depth	Sequencing technology
Dropouts	Various meanings in the literature	Ambiguous
Excess zeros	Various meanings in the literature	Ambiguous
Two-state gene expression model	A model that describes a gene's switching between active and inactive states during transcription	Modeling
Zero inflation	A statistical concept that depends on a specified statistical model	Modeling
Poisson	A statistical model for counts; it requires the count variance to be equal to the count mean	Modeling
Zero-inflated Poisson (ZIP)	A statistical model for counts; it allows for a larger proportion of zeros than Poisson does	Modeling
Negative binomial (NB)	A statistical model for counts; it requires the count variance to be larger than the count mean	Modeling
Zero-inflated negative binomial (ZINB)	A statistical model for counts; it allows for a larger proportion of zeros than NB does	Modeling
Masking scheme	A way to mask a proportion of non-zero counts in a matrix to zeros	Modeling
Differentially expressed (DE) gene	A gene that has statistically significant difference in expression between two conditions (e.g., cell groups)	Modeling
Impute	To change the zero counts in a matrix to non-zero counts	Modeling
Binarize	To change the non-zero counts in a matrix to ones	Modeling

Table 2 Clarification of zero-related terminology

In the current scRNA-seq literature, much ambiguity exists in the use of terms including "dropouts", "excess zeros", and "zero inflation" to describe the prevalence of zeros in scRNA-seq data [94]. We clarify the three terms by summarizing their various uses in the scRNA-seq field to facilitate our discussion.

Dropout or **dropouts** are widely used regarding the prevalence of zeros in scRNA-seq data. It was first introduced in the SCDE method paper: "dropout describes zero gene expression for the genes that show moderate or high expressions in only a proportion of cells [38]". Hence, dropouts, as a data-driven concept, are not equivalent to either biological or non-biological zeros. Nevertheless, the use of "dropouts" in later papers became inconsistent and confusing: most papers meant non-biological zeros [20, 36, 40, 52, 55, 95, 96]; some meant non-biological zeros and low expression measurements [45, 97]; some meant all zeros [46, 47, 98]. In addition, "dropout" was often used as an adjective to mean the existence of many zeros [99]. Such inconsistent uses of "dropouts" are emphasized in a recent work [94]. To avoid possible confusion, we will not use "dropout" or "dropouts" in the following text.

Excess zeros are used in various ways: some papers referred to the larger proportion of zeros in scRNA-seq data than in bulk RNA-seq data [40]; some meant non-biological zeros [45, 96]; some meant the additional zeros that cannot be explained by the negative binomial (NB) model [97]. To avoid confusion, we will not use "excess zeros" in the following text.

Zero inflation, unlike the first two terms, is a statistical concept that depends on a specified model, i.e., a count distribution such as the Poisson distribution and the NB distribution [95]. It means the proportion of zeros that exceeds what is expected under the specified model [40]. We will use "zero inflation" in the following discussion because its definition has no ambiguity.

and non-biological zeros are hardly distinguishable in scRNA-seq data without biological knowledge or spike-in control (see the "Future directions" section).

Biological zeros in scRNA-seq data

A biological zero is defined as the true absence of a gene's transcripts or messenger RNAs (mRNAs) in a cell [67]. Biological zeros occur for two reasons (Fig. 1b). First, many genes are unexpressed in a cell (e.g., gene 1 in Fig. 1b), and cells of distinct types have different genes expressed—a fact that results in the diversity of cell types [68, 69]. Second, many genes undergo a bursty process of transcription (i.e., mRNA synthesis); that is, these genes are not transcribed constantly but intermittently, a well-known phenomenon in gene regulation [38, 39, 46, 70–72]. Specifically, in eukaryotic cells, transcription is initiated by the binding of specific transcription factors (TFs) and RNA polymerase to the promoter of a gene [73–75]. Due to the stochasticity of TF binding, a gene switches between active and inactive states, and its transcription only occurs during the active state [76]. Hence, systems biologists have used a two-state gene expression model to describe how the rates of three processes—active/inactive state switching, transcription, and mRNA degradation—jointly determine the distribution of a gene's mRNA copy numbers, i.e., expression levels, in cells of the same type [76–78]. Figure 2 illustrates the model and provides three example settings of model parameters along with their corresponding gene expression distributions. Depending on the gene's switching rates between the active and inactive states, transcription rate, and degradation rate, the resulting distribution may exhibit a mode near zero, which makes it appear that the gene expresses no mRNA at a particular time, in a large number of cells (e.g., gene 2 in Fig. 1b).

Non-biological zeros in scRNA-seq data

Non-biological zeros reflect the loss of information about truly expressed genes due to the inefficiencies of the technologies employed from sample collection to sequencing. Unlike biological zeros, non-biological zeros refer to the zero expression measurements of genes with transcripts in a cell. There are two types of non-biological zeros [67]: (1) technical zeros, which arise from library-preparation steps before cDNA amplification, and (2)

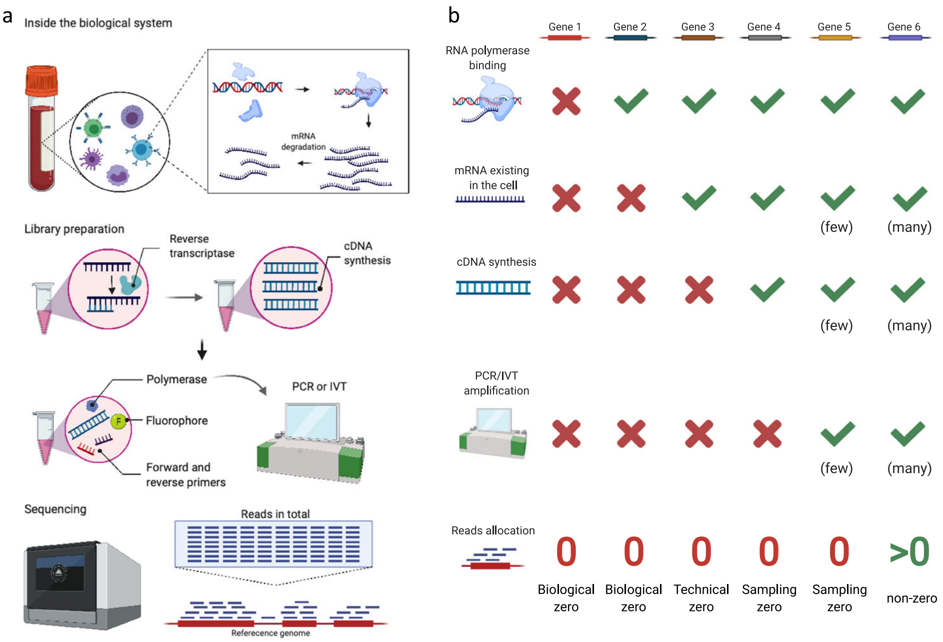


Fig. 1 Sources of zeros in scRNA-seq data. **a** An overview of a scRNA-seq experiment. Biological factors that determine true gene expression levels include transcription and mRNA degradation (top panel). Technical procedures that affect gene expression measurements include cDNA synthesis, PCR or IVT amplification, and sequencing depth (bottom three panels). Finally, every gene's expression measurement in each cell is defined as the number of reads or UMIs mapped to that gene in that cell. **b** How the biological factors and technical procedures in (**a**) lead to biological, technical, and sampling zeros in scRNA-seq data. Red crosses indicate occurrences of zeros, while green checkmarks indicate otherwise. Biological zeros arise from two scenarios: no transcription (gene 1) or no mRNA due to faster mRNA degradation than transcription (gene 2). If a gene has mRNAs in a cell, but its mRNAs are not captured by cDNA synthesis, the gene's zero expression measurement is called a technical zero (gene 3). If a gene has cDNAs in the sequencing library, but its cDNAs are too few to be captured by sequencing, the gene's zero expression measurement is called a sampling zero. Sampling zeros occur for two reasons: a gene's cDNAs have few copies because they are not amplified by PCR or IVT (gene 4), or a gene's mRNA copy number is too small so that its cDNAs still have few copies after amplification (gene 5). If the factors and procedures above do not result in few cDNAs of a gene in the sequencing library, the gene would have a non-zero measurement (gene 6). The figure is created with https://biorender.com/

sampling zeros, which result from inefficient amplification and/or a limited sequencing depth.

One cause of technical zeros is the imperfect mRNA capture efficiency in the reverse transcription (RT) step from mRNA to cDNA. The efficiency has a considerable variation across protocols and may be as low as 20% [79], depending on multiple experimental parameters [80]. The efficiency may even differ between mRNA transcripts. For example, if an mRNA transcript has an intricate secondary structure or is bound to proteins, it would not be reversely transcribed to cDNA efficiently [10, 33, 38]. In summary, if a gene's mRNA transcripts in a cell are not converted into cDNA molecules (cDNAs), the gene would falsely appear as non-expressed in that cell in the sequencing library, resulting in a technical zero in scRNA-seq data (e.g., gene 3 in Fig. 1b).

The other type of non-biological zeros, sampling zeros, occurs due to a constraint on the total number of reads sequenced, i.e., the sequencing depth [64, 81], which is determined by the experimental budget and sequencing machine. During sequencing, cDNAs are randomly captured ("sampled") and sequenced into reads. Hence, a gene with fewer cDNAs is more likely to be undetected due to this random sampling. If undetected, the

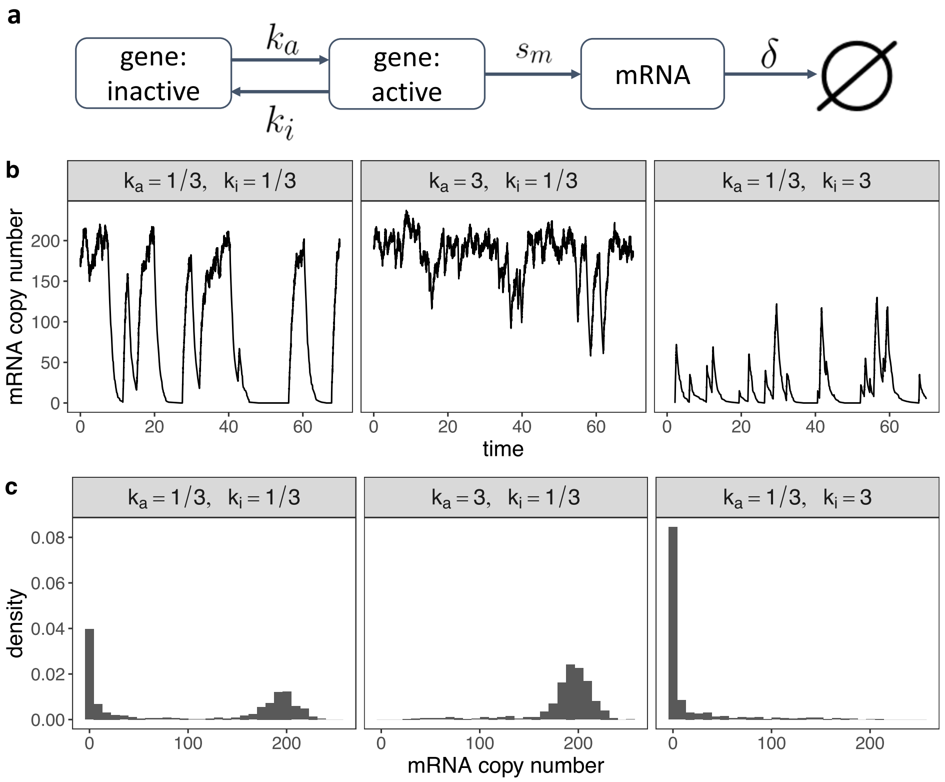


Fig. 2 A two-state stochastic model of the expression levels of one gene. **a** A diagram of the two-state gene expression model [76–78], where a gene stochastically switches from an inactive state to an active state at rate k_a and from an active state to an inactive state at rate k_i. The gene transcribes mRNA at rate s_m only when it is in the active state. The transcribed mRNA then degrades at rate δ. **b** Given $s_m = 200$ and $\delta = 1$, the effects of k_a and k_i on the temporal dynamics of the gene's mRNA copy number. Three example values of k_a and k_i are provided. Left: when both k_a and k_i are small, the mRNA copy number switches between small and large values. Middle: when k_a is much larger than k_i, the mRNA copy number remains large most of the time. Right: when k_a is much smaller than k_i, the mRNA copy number remains small most of the time. **c** Distributions of the gene's mRNA copy number (across cells) corresponding to the three example settings in (**b**). Left: when the gene's mRNA copy number switches between small and large values, the resulting distribution is bimodal with two modes at zero and around s_m/δ. Middle: when the gene's mRNA copy number is large most of the time, the resulting distribution has a single mode around s_m/δ. Right: when the gene's mRNA copy number is small most of the time, the resulting distribution has a single mode at zero. In summary, when k_a is small, the gene is expected to have biological zeros in cells with non-negligible probability

gene's resulting zero read count is a "sampling zero". There are two reasons why a gene (in a cell) may have too few cDNAs in the sequencing library: too few cDNAs before amplification and inefficient cDNA amplification. Below we explain why cDNA amplification may cause some genes to have a disproportionally low cDNA copy number in the sequencing library (see Fig. 3).

The cDNA amplification step is essential for scRNA-seq, as it increases the number of cDNA copies of a gene so that the gene is more likely to be detected by sequencing. Polymerase chain reaction (PCR) [82] is the most widely used amplification procedure. However, PCR amplification is non-linear; thus, the ratio between the copy numbers of two differentially expressed genes is artificially distorted by PCR, i.e., a ratio greater (or smaller) than one becomes even larger (or smaller) after PCR. As a remedy, in vitro transcription (IVT) has been developed for linear amplification [83]. However, compared with PCR, IVT requires more input cDNAs to ensure successful amplification; thus, PCR is

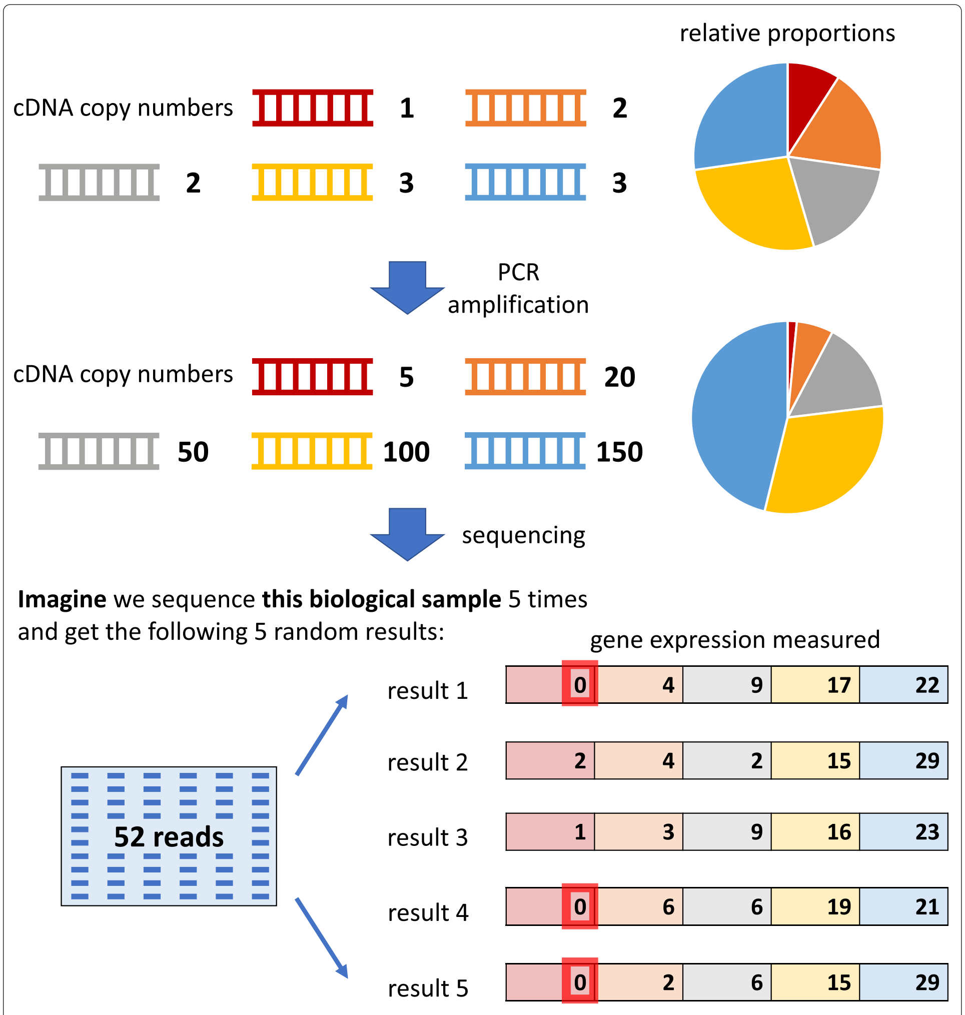


Fig. 3 A toy example showing how the PCR amplification may result in sampling zeros. Five genes have their cDNAs amplified by PCR. After the non-linear amplification, their relative proportions change. If the sequencing depth is limited to 52 reads, the first gene has sampling zeros in three out of five hypothetical sequencing experiments

still the dominant amplification procedure for scRNA-seq [84]. Though indispensable, cDNA amplification is known to introduce biases into cDNA copy numbers because the amplification efficiency depends on cDNA sequence and structure [85, 86]. For example, GC-rich cDNA sequences are harder to be amplified [67]. The amplification efficiency also depends on the design of cell barcodes, adapters, and primers; overlaps or complementarity of barcode, adapter, and primer sequences would induce cDNA secondary structures and thus reduce the amplification efficiency [87, 88]. Moreover, cDNA copy number biases would accumulate as the number of amplification cycles increases [85, 89]; that is, the more cycles, the larger the difference of two genes (with different amplification efficiency) in cDNA copy numbers [90, 91]. Since different scRNA-seq experiments may use different numbers of amplification cycles (e.g., 18 cycles in a Smart-seq2 experiment [92] and 14 cycles in a 10x Genomics experiment [93]), cDNA copy number biases differ among scRNA-seq datasets. In addition, the non-linear amplification nature of PCR would exaggerate the expression level differences between lowly expressed and highly expressed genes. Altogether, due to amplification biases, cDNA copy numbers in a sequencing library may not reflect cDNAs' actual proportions before amplification. As a result, the genes with small cDNA proportions in the sequencing library are likely to be missed by sequencing and thus result in sampling zeros (e.g., gene 4 suffering from inefficient amplification and gene 5 having too few cDNAs in Fig. 1b).

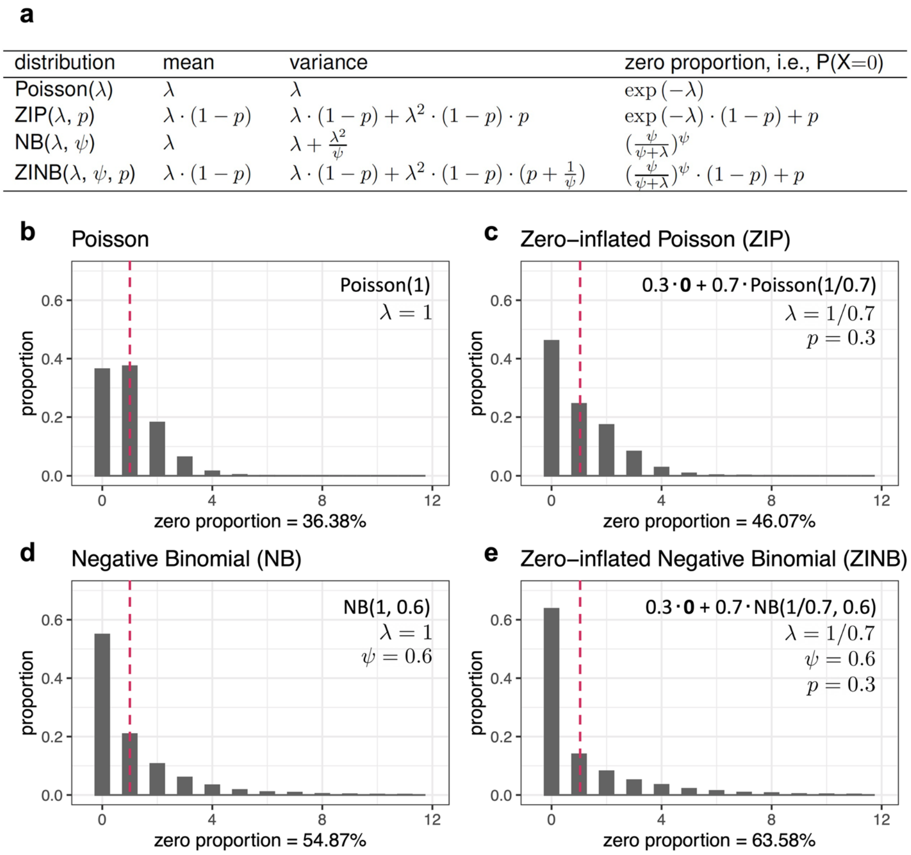
Debate on zero-inflated modeling of scRNA-seq data

Since the advent of scRNA-seq, zero-inflated models have been widely used in bioinformatics tools on the observed scRNA-seq count data [36, 38, 46, 100]. Zero-inflated models are mixture probabilistic models with two components: a point mass at zero and a common distribution, including the Poisson and NB distributions for read or UMI counts and the normal distribution for log-transformed read or UMI counts. More recently, however, researchers have found that UMI counts are not zero-inflated when compared with the Poisson or NB distribution [6, 65, 67, 93, 94]. In particular, Kim et al. provided comprehensive evidence that zero UMI counts can be accounted for by either the NB distribution or even simply the Poisson distribution [65].

The use of UMIs in scRNA-seq can correct the amplification biases in non-zero gene expression measurements [101]; that is, UMIs can be used to identify and remove reads from cDNA duplicates that are results of amplification, and thus some non-zero gene expression measurements would be reduced. However, UMIs cannot help recover sampling zeros, whose corresponding cDNA copy numbers stay unknown despite the use of UMIs [100]. In fact, UMIs cannot reduce any zeros, including biological and non-biological ones. The change of modeling choice—from zero-inflated models for non-UMI-based data to non-zero-inflated models for UMI-based data—indicates that whether or not to use zero-inflated models has nothing to do with the prevalence of zeros. In other words, the modeling choice is a statistical consideration and says nothing about the proportions of zeros or the distinction between biological and non-biological zeros.

Four count distributions—Poisson, zero-inflated Poisson (ZIP), NB, and zero-inflated NB (ZINB)—have been widely used to model a single gene's read or UMI counts across cells in scRNA-seq data. In fact, the former three models are special cases of the ZINB model (Fig. 4a). Poisson only has one parameter (λ) equal to both mean and variance (Fig. 4b). Compared with Poisson, ZIP has one more zero-inflation parameter (p) to indicate the proportion of additional zeros that do not come from Poisson (Fig. 4c); when this zero-inflation parameter is zero, ZIP reduces to Poisson. Also, compared with Poisson, NB has one more dispersion parameter (ψ) that indicates the over-dispersion of variance relative to the mean (i.e., unlike Poisson, NB has variance greater than mean; Fig. 4d); when this dispersion parameter is positive infinity, NB reduces to Poisson. Compared with NB, ZINB has one more zero-inflation parameter (p) to indicate the proportion of additional zeros that do not come from NB (Fig. 4e); when this zero-inflation parameter is zero, ZINB reduces to NB.

For a fair comparison, we illustrate these four distributions, with example parameters such that they all have the same mean as 1 (Fig. 4b–e). With the same mean, ZIP and NB have more zeros than Poisson does, and ZINB has the most zeros. Between ZIP and NB, which one has more zeros depends on their parameter values, and when they have the same zero proportion, their non-zero distributions are still different. Moreover, when the



a

distribution	mean	variance	zero proportion, i.e., P(X=0)
Poisson(λ)	λ	λ	$\exp(-\lambda)$
ZIP(λ, p)	$\lambda \cdot (1-p)$	$\lambda \cdot (1-p) + \lambda^2 \cdot (1-p) \cdot p$	$\exp(-\lambda) \cdot (1-p) + p$
NB(λ, ψ)	λ	$\lambda + \frac{\lambda^2}{\psi}$	$(\frac{\psi}{\psi+\lambda})^{\psi}$
ZINB(λ, ψ, p)	$\lambda \cdot (1-p)$	$\lambda \cdot (1-p) + \lambda^2 \cdot (1-p) \cdot (p + \frac{1}{\psi})$	$(\frac{\psi}{\psi+\lambda})^{\psi} \cdot (1-p) + p$

Fig. 4 Four count distributions: Poisson, zero-inflated Poisson (ZIP), negative binomial (NB), and zero-inflated negative binomial (ZINB). **a** Parameterization, mean, variance, and zero proportion of each of the four distributions. **b**–**e** Illustration of the probability mass functions of Poisson (**b**), ZIP (**c**), NB (**d**), and ZINB (**e**) distributions that all have mean equal to 1. The horizontal axis indicates each possible value, and the vertical axis indicates the probability of taking each possible value. For each distribution, the parameter values are listed on the top right, and the zero proportion is listed at the bottom

four distributions have the same mean, compared with Poisson and ZIP, NB and ZINB have heavier right tails, i.e., greater probabilities of taking larger values.

Svensson shows that non-zero-inflated distributions can describe the variation in droplet scRNA-seq data by using droplet-based ERCC spike-in data [64]. To evaluate this claim on real scRNA-seq data for both droplet-based and full-length protocols, we perform the similar analysis on three real scRNA-seq PBMC datasets. More specifically, we fit the above four count distributions—two zero-inflated (ZIP and ZINB) and two non-zero-inflated (Poisson and NB)—to a non-UMI-based dataset generated by Smart-seq2 and two UMI-based datasets generated by 10x Genomics and Drop-seq. These three datasets are ideal for studying how the modeling choice depends on the experimental protocol, as they were generated by a benchmark study [43] that applied multiple scRNA-seq protocols to measure peripheral blood mononuclear cells (PBMCs) from the same batch, and the benchmark study labeled cells using the same cell types and curated genes. We first compare the three datasets in terms of their distributions of cell library size (i.e., the total number of reads or UMIs in each cell), numbers of cells, and distributions of the number of genes detected per cell. Figure 5a–c show that, compared with the two UMI-based datasets, the Smart-seq2 (non-UMI-based) dataset has larger cell library sizes,

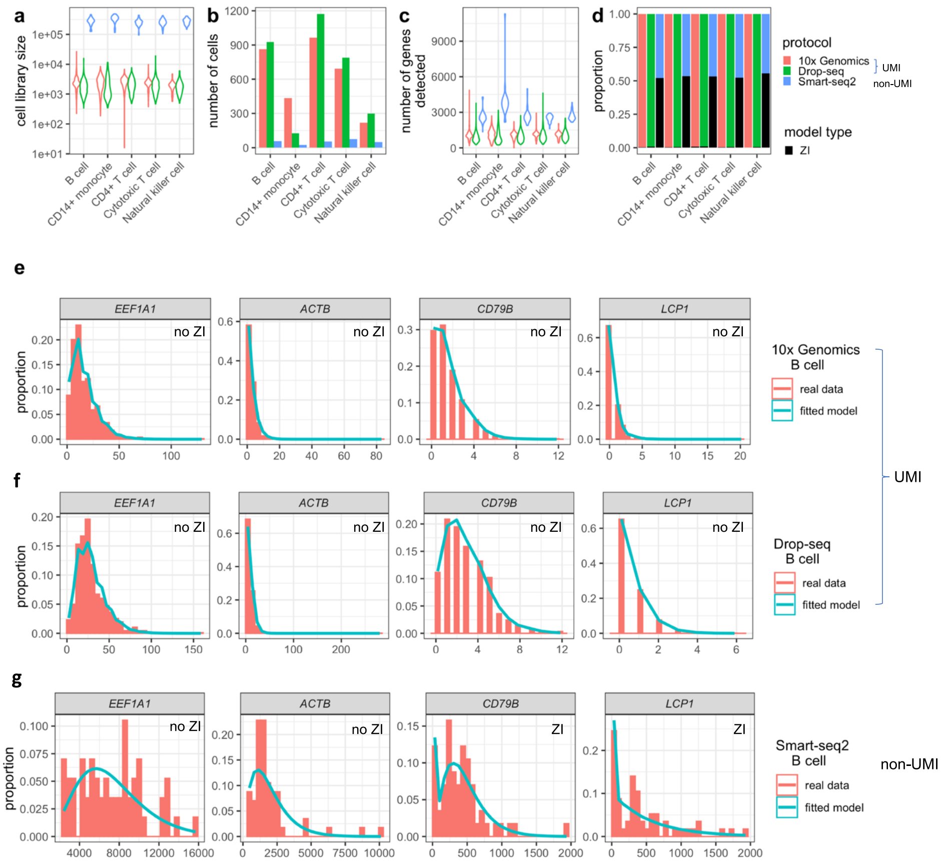


Fig. 5 Statistical modeling of 10x Genomics, Drop-seq, and Smart-seq2 data for the same PMBC sample. 10x Genomics and Drop-seq data are UMI-based, while Smart-seq2 data are non-UMI-based. **a** Violin plots showing the distribution of cell library size for each of five PMBC cell types measured by each protocol [43]. For the UMI-based protocols (10x Genomics and Drop-seq) and Smart-seq2, a cell's library size is defined as the total number of UMIs and reads, respectively, in that cell. **b** Barplots showing the number of cells detected for each cell type by each protocol. **c** Violin plots showing the distribution of the number of genes detected per cell. **d** Barplots showing the proportions of genes for which zero-inflated (ZI) models are chosen (black) and genes for which non-zero-inflated models are chosen (non-black). The model selection is done by likelihood ratio tests. **e** Four example genes' UMI count distributions in B cells measured by 10x Genomics. The observed count distributions are shown in histograms. Non-zero-inflated (no ZI) models are chosen, and the fitted model distributions are shown in cyan curves. **f** The same four example genes' UMI count distributions counts in B cells measured by Drop-seq. **g** The same four example genes' read count distributions in B cells measured by Smart-seq2. ZI models are chosen for two genes

fewer cells, and more genes detected—a phenomenon consistent across the five cell types (B cells, CD14+ monocytes, CD4+ T cells, cytotoxic T cells, and natural killer cells).

Next, for each gene in each dataset, we fit the four distributions to its read or UMI counts in cells of each type, and we choose its distribution among the four distributions by likelihood ratio tests (see [102] for detail). The rationale is to choose the least complex distribution that fits the data well. Figure 5d shows that non-zero-inflated distributions (Poisson and NB) are chosen for almost all genes in the 10x Genomics and Drop-seq datasets, while zero-inflated distributions (ZIP and ZINB) are chosen for about half of the genes in the Smart-seq2 dataset. This result is consistent with the recent advocate for not using zero-inflated models for UMI-based data [64, 94], and it suggests that zero-inflated modeling is still useful for Smart-seq2 data. For illustration purposes, in Fig. 5e–g, we plot the read or UMI count distributions for four genes (*EEF1A1*, *ACTB*, *CD79B*, and *LCP1*) in B cells in these three datasets, and we also plot the fitted chosen distribution for each gene. Specifically, non-zero-inflated distributions are chosen for all the four genes in the UMI-based datasets, while zero-inflated distributions are chosen for *CD79B* and *LCP1* in the Smart-seq2 dataset. Our results show that the same gene's expression distribution under the same biological condition may be described by different statistical models for data generated by different protocols, confirming that zero inflation provides no direct information on biological zeros, whose existence does not depend on protocols (Fig. 1b).

How non-biological zeros affect scRNA-seq data analysis

To evaluate the effects of non-biological zeros on scRNA-seq data analysis, such as cell clustering and DE gene identification, we need access to true cell types and true DE genes. Hence, we use scDesign2 [102], a probabilistic, flexible simulator we developed to generate realistic scRNA-seq count data from any protocol with gene correlations captured. First, we train scDesign2 on the three benchmark PBMC datasets (10x Genomics, Drop-seq, and Smart-seq2) [43], which all contain the same five cell types (B cells, CD14+ monocytes, CD4+ T cells, cytotoxic T cells, and natural killer cells) and are used in Fig. 5. Second, we simulate the corresponding non-zero-inflated synthetic datasets, one for each protocol, in the form of gene-by-cell count matrices. In detail, after the first training step, every gene in each cell type has a fitted count distribution (Poisson, ZIP, NB, or ZINB) by scDesign2; in the second simulation step, we generate read or UMI counts for every gene in each cell type from the non-zero-inflation component (Poisson or NB). Note that we set the number of synthetic cells generated by scDesign2 equal to the number of real cells for each cell type. Hence, for each gene, this simulation procedure removes the statistical zero inflation, which we define in the last section, and provides the gene's expected expression level in each cell type (as the mean of its non-zero-inflation component).

We use scDesign2 [102] as the simulator because we desire synthetic cells that preserve real genes and gene-gene correlations observed in real data. The reason is that real genes are the targets of DE analysis, and synthetic cells with realistic gene-gene correlations are necessary for evaluating cell clustering and dimension reduction. As discussed in the scDesign2 paper [102], simulators such as SymSim [103] and Splatter [68] do not preserve real genes, and Sergio [104] requires an additional input of a gene regulatory network and cannot preserve the observed gene-gene correlations in real data. We note that we do not aim to benchmark simulators in this work, so we choose scDesign2, which preserves genes and gene-gene correlations and allows us to generate non-zero-inflated data, making it easy for us to introduce non-biological zeros using various masking schemes.

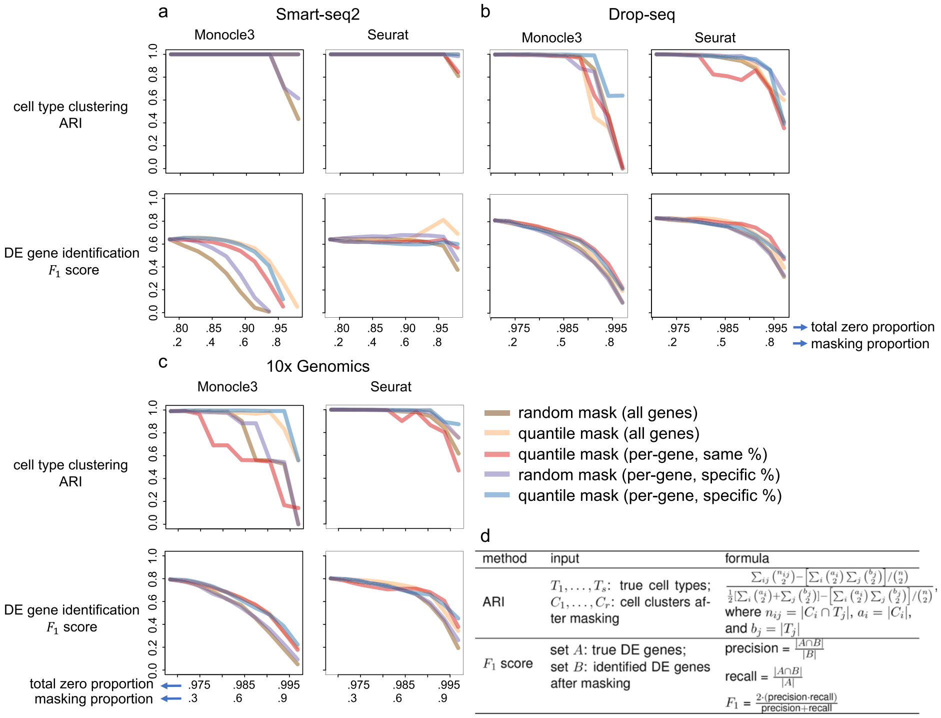
To simulate scRNA-seq data with known DE genes, we first use scDesign2 to fit probabilistic models—selected from Poisson, NB, ZIP, and ZINB models—to each cell type in each real PBMC dataset generated by Smart-seq2, Drop-seq, or 10x Genomics. Then we focus on two cell types, CD4+ T cells and Cytotoxic T cells, and define the true DE genes as the top 1500 genes that have the largest estimated mean differences (in the fitted probabilistic models) between the two cell types. We consider the remaining genes as true non-DE genes, and we set the mean parameter of each true non-DE gene to be the same for the two cell types (by averaging the gene's estimated mean parameters in its two fitted models for the two cell types). Finally, we use scDesign2 to generate synthetic scRNA-seq data without zero inflation: for every true DE gene, we generate its synthetic counts from its two fitted models (with zero inflated components removed if existent) for the two cell types; for every true non-DE gene, we generate its synthetic counts from two altered models (with the same averaged mean, zero inflation components removed if existent, and NB's dispersion parameters kept as they are estimated in the two fitted models).

Based on the three non-zero-inflated synthetic datasets (10x Genomics, Drop-seq, and Smart-seq2), we define the positive controls for two typical analyses: cell clustering and DE gene identification, which are ubiquitous in scRNA-seq data analysis pipelines. For cell clustering, the positive controls are provided by scDesign2 as the cell types from which it generates synthetic cells. For DE gene identification, the positive controls are provided by scDesign2 as the genes whose expected expression levels differ between cell types.

Using each of the five masking schemes (see Additional file 1), we introduce a varying number of non-biological zeros, corresponding to masking proportions $p = 0.1, \ldots, 0.9$, into the three synthetic datasets corresponding to 10x Genomics, Drop-seq, and Smart-seq2 protocols, creating three suites of zero-inflated datasets, one suite per protocol. Note that each suite contains one non-zero-inflated dataset and $45 = 9$ (# of masking proportions) × 5 (# of masking schemes) zero-inflated datasets. Then we apply Monocle3 (R package version `0.2.3.0`) [28] and Seurat (R package version `3.2.1`) [13], two popular multi-functional software packages, to the three suites of datasets. We use the two packages to perform cell clustering and DE gene identification, and we evaluate the analysis results based on our previously defined positive controls. Figure 6a–c summarizes how the accuracy of the two analyses deteriorates as the masking proportion increases under each masking scheme and for each protocol.

The clustering results (top rows in Fig. 6a–c) show that the clustering accuracy (measured by the adjusted rand index; Fig. 6d) is robust to the introduction of non-biological zeros up to the masking proportion $p = 0.6$ (i.e., 60% non-zero counts are masked as zeros) for most masking schemes. Compared with Monocle3, Seurat is more robust to non-biological zeros under all the five masking schemes. Among all schemes, the two schemes that assume (1) dependence between masking and count values and (2) gene-specific masking proportions—`quantile mask (all genes)` and `quantile mask (per-gene, specific %)` (Additional file 1: Figure S1)—have the least deteriorating effects on cell clustering. This result is reasonable as these two schemes tend to mask low counts to zeros so that the relative order of gene expression counts (from low to high) is better preserved than by the other three schemes. A recent article argues that zeros in scRNA-seq data carry biological meanings and should be embraced, and its argument is based on the assumption that most zeros correspond to low expression levels [66], an assumption aligned with these two masking schemes. Finally, among the three protocols, clustering on Smart-seq2 data is most robust to non-biological zeros, likely because Smart-seq2 data contain fewer zeros than the two UMI-based protocols' data do. It is worth noting that, between the two UMI-based protocols, clustering accuracy is better on 10x Genomics data than Drop-seq data.

The DE gene identification results (bottom rows in Fig. 6a–c) show that the F_1 scores (at 5% false discovery rate; Fig. 6d) are robust to non-biological zeros for Seurat, but not as much for Monocle3. The reason is that Seurat uses MAST [36], a method built upon a zero-inflated model, for DE gene identification, while Monocle3 uses non-zero-inflated models (including Poisson, quasi Poisson, and NB) that cannot account for additional non-biological zeros. Among the five masking schemes, the two random schemes that assume independence between masking and count values—`random mask`



method	input	formula
ARI	$T_1, \ldots, T_s$: true cell types; $C_1, \ldots, C_r$: cell clusters after masking	$\frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]/\binom{n}{2}}$, where $n_{ij} = \lvert C_i \cap T_j \rvert$, $a_i = \lvert C_i \rvert$, and $b_j = \lvert T_j \rvert$

F_1 score	set A: true DE genes; set B: identified DE genes after masking	precision = $\frac{\lvert A \cap B \rvert}{\lvert B \rvert}$
		recall = $\frac{\lvert A \cap B \rvert}{\lvert A \rvert}$
		F_1 = $\frac{2 \cdot (\text{precision} \cdot \text{recall})}{\text{precision} + \text{recall}}$

Fig. 6 Effects of non-biological zeros on cell clustering and DE gene identification. We introduce a varying number of non-biological zeros, which correspond to masking proportions 0.1–0.9, into the simulated **a** Smart-seq2, **b** Drop-seq, and **c** 10x Genomics datasets using five masking schemes. The horizontal axes show (top) the total zero proportion (including the zeros before masking and the non-biological zeros introduced by masking) and (bottom) the masking proportion (i.e., the proportion of non-zero counts masked by a masking scheme). After introducing non-biological zeros, we apply Monocle 3 and Seurat to each dataset to perform cell clustering and identify DE genes. For the two analyses, we evaluate their accuracy using the adjusted rand index (ARI) and F_1 score (given the false discovery rate 5%), respectively. **d** Technical definitions of the ARI and F_1 score

`(all genes)` and `random mask (per-gene, specific %)` (Additional file 1: Figure S1)—have the most deteriorating effects on DE gene identification. This result is reasonable as these two schemes may mask high counts to zeros, so they disrupt every gene's count distribution more than the other three schemes do. Interestingly, although `quantile mask (per-gene, same %)` is unlikely a realistic generation mechanism of non-biological zeros as it masks the same proportion of non-zero counts for every gene, we observe that Seurat has robust F_1 scores as non-biological zeros are introduced by this scheme. This seemingly unexpected result reflects that zero-inflated models are robust for DE gene identification under quantile masking, even though the masking proportion may not be reasonable. Finally, regarding the three protocols, Seurat has better F_1 scores for Smart-seq2 data than Monocle3 does, a reasonable result given the observed zero-inflation in Smart-seq2 data (Fig. 5d). For the two UMI-based protocols, Monocle3 and Seurat have comparable performance in terms of F_1 scores. We have also observed that the DE analysis results for UMI-based data are better than for non-UMI-based data. One possible reason is the larger sample sizes (larger numbers of cells) in Drop-seq and 10x data that increase the power in statistical testing. To supplement the F_1 scores, we show the corresponding precision and recall rates in Additional file 1: Figure S2. It is worth noting that although the false discovery rate is set to 5%, the precision rates of

Monocle3 and Seurat are sometimes far below the expected precision 95%, which is equal to one minus the false discovery rate. This phenomenon calls for better false discovery rate control in scRNA-seq DE analysis [105]. In addition, compared to Seurat, Monocle3 shows a greater fluctuation in both precision and recall as the masking proportion increases for Smart-seq2 data.

In summary, compared with DE gene identification, cell clustering is more robust to non-biological zeros. This result suggests that the sparsity in scRNA-seq data affects gene-level analyses more than cell-level analyses because the latter jointly uses all genes' expression levels. Overall, Seurat is more robust than Monocle3 is to non-biological zeros for both analyses. For cell clustering, Seurat has better accuracy regardless of protocols. For DE gene identification, Seurat is preferable for Smart-seq2 data, while Monocle3 has better accuracy for UMI-based data.

It is worth noting that many imputation methods evaluate their imputation accuracy based on only the `random mask (all genes)` scheme [49, 62, 106]. Our results indicate that non-biological zeros introduced by different masking schemes have different effects on cell clustering and DE gene identification, and quantile masking may be more realistic given previous reports that genes with lower expression values have more zeros than genes with higher expression [39, 46]. Hence, we urge that quantile masking schemes be considered in the future evaluation of computational methods that deal with non-biological zeros.

Input data: observed vs. imputed vs. binarized counts

Current scRNA-seq data analysis typically takes three types of input data: observed, imputed, and binarized counts. Researchers use imputed and binarized counts to deal with the vast proportion of zeros. Although log-transformed counts are often used as input data, this practice is under controversy [93, 107, 108] and is not the focus of our discussion. Here we summarize the advantages, disadvantages, and suitable users (bioinformatics tool developers vs. users) of each input data type.

Direct modeling of observed counts is the most common practice for bioinformatics tool developers [13, 28, 30–33, 36, 109]. An obvious advantage of direct modeling is that observed counts are not biased by any data pre-processing steps. Hence, observed counts are the preferred input data type for most tool developers. However, unlike tool developers, tool users need to apply existing bioinformatics tools to scRNA-seq data. If the observed counts do not work well with existing tools, for practical reasons, tool users may consider data pre-processing steps such as imputation and binarization so that existing tools can output reasonable analysis results.

Since the sparsity in scRNA-seq counts has posed a great hurdle for many existing tools, imputation has been proposed as a practical data pre-processing step, and many imputation methods have been developed [20, 44, 45, 47–63]. Of course, imputation has the risk of biasing data, leading to false signals [99] or diminished biological variation [45, 63, 110]. For example, Chen et al. pointed out that a major drawback of scRNA-seq imputation is diminished gene expression variability across cells after imputation; they argued that it would be important to quantify the expression variability before and after imputation, in addition to evaluating the mean expression prediction (by, for example, comparing it to the gene expression measurement in the same cell type in a separate bulk dataset) [63]. However, imputation has two practical advantages for single-cell biologists who are mostly tool users. First, many imputation methods have shown that their imputed counts, in which many zeros in the observed counts become non-zeros, agree better with biological knowledge and/or biologists' expectations. For example, the effectiveness of imputation has been supported by evidence that scRNA-seq data after imputation agree better with bulk RNA-seq data or single-cell RNA fluorescence in situ hybridization (FISH) data [48, 62, 111]. Second, imputation builds a bridge that connects sparse scRNA-seq data to many powerful tools designed for non-sparse data. For example, DESeq2 [35] and edgeR [31] are two popular DE gene identification methods for bulk RNA-seq data; however, they are not directly applicable to scRNA-seq data because their models do not account for data sparsity. Hence, if tool users cannot find a DE gene identification method that works well for their scRNA-seq data, they may consider reducing zeros by imputation methods to make DESeq2 or edgeR applicable [60, 61, 63], conditional on verified false discovery rate control [105, 112].

Moreover, a recent article by Qiu provides a new perspective by proposing to use only binarized counts (with all non-zero counts truncated as ones) for cell clustering [66]. It argues that, by removing the magnitudes of non-zero counts, binarization alleviates the need for normalizing individual cells' sequencing depths. Further, its key message is that zeros are biologically meaningful because binarized counts can lead to reasonable cell clustering results. Other works also suggest that binarized counts can serve as useful data, in addition to observed counts, and be incorporated into scRNA-seq data modeling and analysis [23, 113]. Although binarized counts eliminate the expression differences between highly and lowly expressed genes, they highlight the co-expression patterns of genes, i.e., whether two genes are co-expressed in a cell, which have been used in marker gene selection [12] and gene network construction [114–116]. However, it remains unclear whether binarized counts can replace observed counts in scRNA-seq data analysis. Our intuition says that the answer is unlikely yes for all analyses because the magnitudes of non-zero counts reflect expression levels of genes in each cell. Qiu uses binarized counts to deal with cell clustering, a cell-level analysis [66]. For gene-level analyses such as DE gene identification, binarized counts are unlikely better than observed counts. For example, if a gene has similar percentages of zero counts in two cell types, but its non-zero counts are much larger in one cell type than the other, then this gene should be identified as DE using observed counts, but it would be missed as DE using binarized counts. In the previous section "How non-biological zeros affect scRNA-seq data analysis," we have compared the effects of non-biological zeros on clustering and DE gene analysis. For tool developers, it would be beneficial to consider using binarized counts in addition to observed counts for developing new analysis tools. For tool users, binarized counts can be used for exploratory data analysis because several efficient computational tools are applicable to binary counts only, e.g., scalable probabilistic principal component analysis [117].

We further evaluate the effects of the three input data types (observed, binarized, and imputed counts) on three popular downstream analyses: cell clustering, cell dimension reduction (two-dimensional visualization), and DE gene identification. To obtain the imputed counts, we use three popular imputation methods: scImpute, MAGIC, and SAVER, which demonstrate good performance in a recent study that benchmarked 18 imputation methods [118]. We note that our goal here is not to benchmark the existing > 70 imputation methods (https://www.scrna-tools.org/tools?sort=name&cats=Imputation) but to demonstrate the importance of considering various zero generation processes (i.e., masking schemes) to achieve fair benchmarking of computational methods. For readers interested in evaluating other imputation methods, we have released our code for the five masking schemes (Additional file 1) on GitHub (https://github.com/ruochenj/Five_masking_schemes).

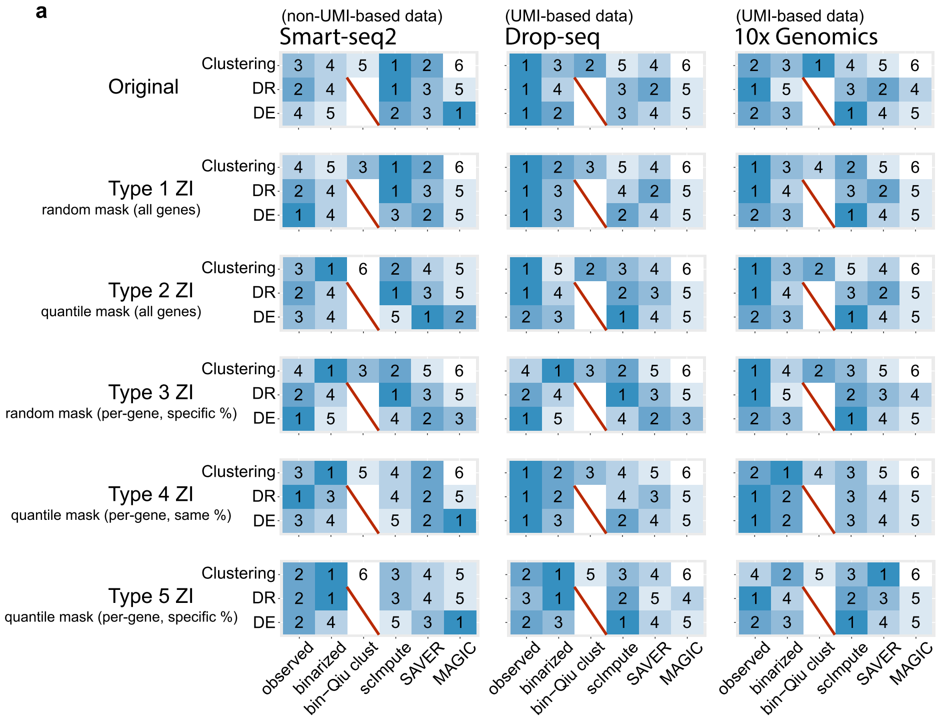
To benchmark cell clustering and dimension reduction results, we use the three real scRNA-seq PBMC datasets with labeled cell types [43]—one non-UMI-based dataset generated by Smart-seq2 and two UMI-based datasets generated by 10x Genomics and Drop-seq—which we have used in the previous section to evaluate the effects of non-biological zeros. To benchmark DE gene identification results, we generate synthetic datasets containing pre-defined true DE genes by scDesign2 [102] from two cell types (CD4+ T cells and cytotoxic T cells) in the three real datasets. (The definition of DE genes is described in the previous section.)

For cell clustering, we use two algorithms: Qiu's algorithm designed specifically for binarized counts [66] and the Louvain algorithm (implemented in Seurat). For all three input data types, we use the Louvain algorithm to cluster cells; for binarized counts only, we also use Qiu's algorithm. Based on the cell type labels provided in all three datasets, we calculate the ARI as a measure of clustering accuracy (Additional file 1: Figure S3; top row).

For cell dimension reduction (two-dimensional visualization), we perform PCA, t-SNE, and UMAP (implemented in Seurat) on the observed and imputed counts, and then we use the average Silhouette score to evaluate how well the labeled cell types are separated in the two-dimensional space (Additional file 1: Figure S4). We observe that UMAP has the overall best performance and thus decide to use UMAP to evaluate the impacts of imputation methods and binarization on the dimension reduction analysis (Additional file 1: Figure S5). Additional file 1: Figures S6, S7, and S8 show UMAP's two-dimensional reduction results of Smart-seq2, Drop-seq, and 10x Genomics data, respectively.

For DE gene analysis, we consider two DE methods. For all three inpute data types, we use MAST (implemented in Seurat) to perform DE gene identification; for binarized counts only, we also apply a two-sample proportion test to the binarized data. At a 5% false discovery rate, we use precision (Additional file 1: Figure S9), recall (Additional file 1: Figure S10), and F_1 score (Additional file 1: Figure S11) to evaluate the identification results.

Figure 7a, b summarizes the relative performance of the three input data types for the three protocols (Smart-seq2, Drop-seq, and 10x Genomics) in the three downstream analyses. In terms of cell clustering, for non-UMI-based Smart-seq2 data, the Louvain algorithm has better performance on scImpute and SAVER's imputed counts than on the observed or binarized counts; for UMI-based Drop-seq and 10x Genomics data, the Louvain algorithm on the observed counts and Qiu's algorithm on the binarized counts have comparable performance and outperform the Louvain algorithm applied to other input data types, suggesting that imputation does not improve the clustering of UMI-based data. Notably, Qiu's algorithm only works well for binarized counts of UMI-based data, likely due to its special design. In terms of cell dimension reduction, scImpute's imputed counts work the best for non-UMI-based data; the observed counts have the best performance for UMI-based data; binarized counts and MAGIC's imputed counts have poor performance for both non-UMI-based and UMI-based data. In terms of DE



b

Input data	Smart-seq2	Drop-seq	10x Genomics	Clustering	DR	DE
observed	3.0	**1.0**	**1.7**	**2.3**	**1.5**	2.8
binarized	4.3	3.0	3.7	3.5	4.3	3.8
bin-Qiu clust	NA	NA	NA	3.3	NA	NA
scImpute	**1.3**	3.7	2.7	2.8	2.0	**2.0**
SAVER	2.7	3.3	3.7	3.3	2.5	3.5
MAGIC	4.0	5.3	5.0	6.0	4.8	3.0

Fig. 7 Performance ranks of using observed, binarized, and imputed counts (from three experimental protocols and under five masking schemes) in three downstream analyses. **a** We perform cell clustering (Clustering), cell dimension reduction (DR), and gene differential expression (DE) analysis on the observed, binarized, and imputed counts of Smart-seq2, Drop-seq, and 10x Genomics data. We consider three popular imputation methods: scImpute, SAVER, and MAGIC. In addition to the original data, we use five masking schemes (Type 1 ZI–Type 5 ZI) to introduce 50% non-biological zeros and evaluate the effects on the downstream analyses with different input data. The five masking schemes are `random mask (all genes)`, `quantile mask (all genes)`, `random mask (per-gene, specific %)`, `quantile mask (per-gene, same %)`, and `quantile mask (per-gene, specific %)`, corresponding to type 1 ZI–type 5 ZI, respectively. The six columns correspond to different input data types: observed counts, binarized counts, binarized counts analyzed by the Qiu's clustering algorithm (bin-Qiu clust), imputed counts by scImpute, imputed counts by SAVER, and imputed counts by MAGIC. For cell clustering, except bin-Qiu clust, clustering is conducted by the Louvain clustering algorithm (in Seurat); clustering performance is ranked by the ARI. For cell DR analysis, we apply UMAP (in Seurat) to perform DR and calculate the average Silhouette score (based on known cell types) for each input data type to evaluate the DR performance. For gene DE analysis, we apply the two-sample proportion test to binarized counts and MAST (in Seurat) to observed, binarized, and imputed data to perform DE analysis. To rank the DE performance by the F_1 score (at the 5% false discovery rate), since binarized counts have two DE methods, we compute the rank for the better-performing method in each comparison. In each row of each matrix, rank 1 indicates the best-performing input data type, while rank 6 indicates the worst. **b** On the original data, we compute the average ranks of the six input data types. Columns 2–4 show the average ranks for Smart-seq2 data, Drop-seq data, and 10x Genomics data across the three downstream analysis—cell clustering analysis, cell DR analysis, and gene DE analysis. Columns 5–7 show the weighted averages of the ranks for the three downstream analysis given the three protocols. Weights of 2, 1, 1 are used for Smart-seq2, Drop-seq, and 10x Genomics to ensure that the weights for non-UMI-based and UMI-based data are equal

gene analysis, for non-UMI-based data, all three imputation methods' imputed counts outperform the observed and binarized counts, a result consistent with our previous discussion on the existence of zero-inflation in non-UMI-based data; for UMI-based data, the observed counts and scImpute's imputed counts lead to the best result for Drop-seq and 10x Genomics data, respectively.

Moreover, we evaluate the three input data types in the three downstream analyses after applying the five masking schemes (see Additional file 1) to introducing additional non-biological zeros. Under each masking scheme, we mask 50% of the original nonzero counts as zeros in each of the three original datasets (Smart-seq2, Drop-seq, and 10x Genomics). In terms of cell clustering analysis, for non-UMI-based data, scImpute's imputed counts demonstrate robust performance and stay as a top-performing input data type under the first three masking schemes, including the two random masking schemes; interestingly, by the Louvain algorithm, the binarized counts do not perform well for the original data but become a top-performing input data type under the last four masking schemes, including the three quantile masking schemes. These two results suggest that scImpute's imputation and binarization ameliorate the effects of additional non-biological zeros in complementary ways. For UMI-based data, the observed counts lead to the overall best clustering results under all masking schemes (ranked the 1st in 6 out of 10 protocol-masking scheme combinations), while Qiu's algorithm is not robust to the introduction non-biological zeros by masking schemes. In terms of cell dimension reduction, scImpute's imputed counts are the best input data type for non-UMI-based data (ranked the 1st under 3 out of 5 masking schemes), while the observed counts are the best for UMI-based data (ranked the 1st in 8 out of 10 protocol-masking scheme combinations). In terms of DE gene analysis, for non-UMI-based data, there is no universal winner: the observed counts work the best under random masking schemes, while SAVER and MAGIC's imputed counts work the best under quantile masking schemes. For UMI-based data, scImpute's imputed counts have the best performance (ranked the 1st in 6 out of 10 protocol-masking scheme combinations), followed by the observed counts (ranked the 1st in 4 out of 10 protocol-masking scheme combinations).

In summary, the observed counts work well for UMI-based data and are robust to the introduction of non-biological zeros. As expected, the binarized counts work well under the quantile masking schemes, which largely preserve the ranks of gene expression levels. Qiu's clustering algorithm works well for the binarized counts of UMI-based data but is not robust to the introduction of non-biological zeros. Imputation methods show concrete improvement for non-UMI-based data, but not so much for UMI-based data. This is consistent with the findings by Kim et al. that imputing UMI-based data can introduce unwanted noise and is thus not recommended [65]. Among the imputation methods, scImpute shows the best performance, while MAGIC does not perform well; a likely reason is that the data we use contain discrete cell types instead of continuous cell trajectories. Notably, the performance of imputation methods depends heavily on the masking scheme, demonstrating the importance of considering multiple masking schemes for the development and benchmarking of imputation methods.

Future directions

ScRNA-seq technologies have advanced the revelation of genome-wide gene expression profiles at the cell level. Accordingly, many computational algorithms and statistical models have been developed for analyzing scRNA-seq data. A well-known challenge in scRNA-seq data analysis is the prevalence of zeros, and how to best tackle zeros remains a controversial topic. Modeling and analysis may be performed on observed, imputed, or binarized scRNA-seq counts. However, the relative advantages and disadvantages of these three strategies remain ambiguous. In this article, we attempt to address this controversy by discussing multiple intertwined topics: the biological and non-biological sources of zeros, the relationship between zero prevalence and scRNA-seq technologies, the extent to which zero prevalence affects various analytical tasks, and the three strategies' relative advantages, disadvantages, and suitable users. We benchmark the performance of analytical tasks on observed, binarized, and imputed data with or without introducing additional non-biological zeros using five masking schemes.

The prevalence of biological and non-biological zeros is a mixed result of intrinsic biological nature and complex scRNA-seq experiments. In particular, the generation mechanism of non-biological zeros is protocol dependent. Hence, it is infeasible to distinguish non-biological zeros from biological zeros purely based on observed counts. As a result, existing imputation methods have a glass ceiling if they use only observed counts as input. To better distinguish non-biological zeros from biological zeros, researchers need to utilize spike-in RNA molecules, whose true counts are known (e.g., External RNA Control Consortium spike-ins [119]), to investigate the generation mechanism of non-biological zeros. Such investigation requires consortium efforts such as the work by the Sequencing Quality Control (SEQC-2) consortium [120]. With a better understanding of how the generation of non-biological zeros depends on mRNA sequence features such as GC contents, statistical and mechanistic models may be developed to better distinguish non-biological zeros from biological zeros and thus to improve imputation accuracy.

The prevalence of biological and non-biological zeros is only one of the many obstacles in using scRNA-seq data for scientific discoveries. As scientific discovery is a trial-and-error process, scRNA-seq data analysis is unavoidably multi-step. Hence, bioinformatics tool developers must consider the pre-processing steps applied to input data and the downstream analyses users may perform on output data. Taking the popular Seurat package as an example, many data pre-processing steps are used before DE gene identification. These steps include filtering low-quality genes and cells, data normalization, gene selection, cell dimension reduction, and cell clustering. Hence, if tool developers are not aware of these pre-processing steps, their bioinformatics tools may not fit into the state-of-the-art scRNA-seq data analysis pipelines. Ultimately, the transparency and reproducibility of scRNA-seq data analysis call for a community collaboration between tool developers and users. Towards this goal, every research article, regardless of being tool development or data analysis, should contain a detailed description of each step and the underlying justifications [121].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-022-02601-5.

Additional file 1: Supplementary material. It includes a detailed description of the five masking schemes, which introduce non-biological zeros to a scRNA-seq count matrix, and supplementary figures.

Additional file 2: Review history.

Acknowledgements
The authors would like to thank the comments and feedback from Dr. Yiling Elaine Chen, Mr. Guanao Yan, Mr. Han Cui, Dr. Xinzhou Ge, and other members of the Junction of Statistics and Biology at UCLA (http://jsb.ucla.edu).

Peer review information
Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 2.

Authors' contributions
RJ and JJL designed the research. RJ conducted the research. RJ and TS wrote the computer code. JJL supervised the project. RJ, TS, DS, and JJL discussed the results and wrote the manuscript. All authors read and approved the final manuscript.

Authors' information
Twitter handles: Ruochen Jiang (@Ruochen_Jiang); Tianyi Sun (@TianyiSun17); Dongyuan Song (@SongDongyuan); Jingyi Jessica Li (@jsb_ucla).

Funding
This work was supported by the following grants: National Science Foundation DBI-1846216 and DMS-2113754, NIH/NIGMS R01GM120507 and R35GM140888, Johnson & Johnson WiSTEM2D Award, Sloan Research Fellowship, and UCLA David Geffen School of Medicine W.M. Keck Foundation Junior Faculty Award (to J.J.L.).

Availability of data and materials
The R code (license: Creative Commons Attribution 3.0 United States) for reproducing the results in Figs. 2, 3, 4, 5 and 6 and Additional file 1 is available at https://doi.org/10.5281/zenodo.4393040 [122].

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Statistics, University of California, Los Angeles 90095-1554, CA, USA. [2]Bioinformatics Interdepartmental Ph.D. Program, University of California, Los Angeles 90095-7246, CA, USA. [3]Department of Human Genetics, University of California, Los Angeles 90095-7088, CA, USA. [4]Department of Computational Medicine, University of California, Los Angeles 90095-1766, CA, USA. [5]Department of Biostatistics, University of California, Los Angeles 90095-1772, CA, USA.

Received: 27 September 2021 Accepted: 4 January 2022
Published online: 21 January 2022

References

1. Saliba A-E, Westermann AJ, Gorski SA, Vogel J. Single-cell rna-seq: advances and future challenges. Nucleic Acids Res. 2014;42(14):8845–60.
2. Liu S, Trapnell C. Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Research. 2016;5:5.
3. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell rna-seq data. Nat Rev Genet. 2019;20(5):273–82.
4. Tritschler S, Büttner M, Fischer DS, Lange M, Bergen V, Lickert H, Theis FJ. Concepts and limitations for learning developmental trajectories from single cell genomics. Development. 2019;146(12):dev170506.
5. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5): 1202–14.
6. Salomon R, Kaczorowski D, Valdes-Mora F, Nordon RE, Neild A, Farbehi N, Bartonicek N, Gallego-Ortega D. Droplet-based single cell rnaseq tools: a practical guide. Lab Chip. 2019;19(10):1706–27.
7. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8(1):1–12.
8. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length rna-seq from single cells using smart-seq2. Nat Protoc. 2014;9(1):171–81.
9. Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, Li N, Szpankowski L, Fowler B, Chen P, et al. Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol. 2014;32(10):1053.
10. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A, Teichmann SA. Power analysis of single-cell rna-sequencing experiments. Nat Methods. 2017;14(4):381–7.
11. Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, Huang Y, Wang J. Comparative analysis of droplet-based ultra-high-throughput single-cell rna-seq systems. Mol Cell. 2019a;73(1):130–42.
12. Wang F, Liang S, Kumar T, Navin N, Chen K. Scmarker: ab initio marker selection for single cell transcriptome profiling. PLoS Comput Biol. 2019;15(10):e1007445.

13. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502.
14. Kiselev YV, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, et al. Sc3: consensus clustering of single-cell rna-seq data. Nat Methods. 2017;14(5):483–6.
15. Guo M, Wang H, Potter SS, Whitsett JA, Yan X. Sincera: a pipeline for single-cell rna-seq profiling analysis. PLoS Comput Biol. 2015;11(11):e1004575.
16. Ho Y-J, Anaparthy N, Molik D, Mathew G, Aicher T, Patel A, Hicks J, Hammell MG. Single-cell rna-seq analysis identifies markers of resistance to targeted braf inhibitors in melanoma cell populations. Genome Res. 2018;28(9): 1353–63.
17. Zeisel A, Muñoz-Manchado AB, Codeluppi S, Lönnerberg P, La Manno G, Juréus A, Marques S, Munguba H, He L, Betsholtz C, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. Science. 2015;347(6226):1138–42.
18. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, Park J-E, Stephenson E, Polański K, Goncalves A, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. Nature. 2018;563(7731):347–53.
19. Croft AP, Campos J, Jansen K, Turner JD, Marshall J, Attar M, Savary L, Wehmeyer C, Naylor AJ, Kemble S, et al. Distinct fibroblast subsets drive inflammation and damage in arthritis. Nature. 2019;570(7760):246–51.
20. Lin P, Troup M, Ho JWK. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. Genome Biol. 2017;18(1):59.
21. Sun Z, Wang T, Ke D, Wang X-F, Lafyatis R, Ding Y, Ming H, Chen W. Dimm-sc: a dirichlet mixture model for clustering droplet-based single cell transcriptomic data. Bioinformatics. 2018;34(1):139–46.
22. Yau C, et al. pcareduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics. 2016;17(1): 140.
23. Andrews TS, Hemberg M. M3drop: dropout-based feature selection for scrnaseq. Bioinformatics. 2019;35(16): 2865–7.
24. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381.
25. Ji Zhicheng, Tscan HongkaiJi. Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. Nucleic Acids Res. 2016;44(13):e117—e117.
26. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics. 2018;19(1):477.
27. Qiu X, Qi M, Tang Y, Li W, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods. 2017;14(10):979.
28. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566(7745):496–502.
29. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol. 2019;37(5):547–54.
30. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat methods. 2018;15(4):255.
31. Robinson MD, McCarthy DJ, Smyth GK. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
32. Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, Pawitan Y. Beta-poisson model for single-cell rna-seq data analyses. Bioinformatics. 2016;32(14):2128–35.
33. Miao Z, Ke D, Wang X, Zhang X. Desingle for detecting three types of differential expression in single-cell rna-seq data. Bioinformatics. 2018;34(18):3223–4.
34. Suomi T, Seyednasrollah F, Jaakkola MK, Faux T, Elo LL. Rots: An r package for reproducibility-optimized statistical testing. PLoS Comput Biol. 2017;13(5):e1005562.
35. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol. 2014;15(12):550.
36. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome Biol. 2015;16(1):1–13.
37. Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C. A statistical approach for identifying differential distributions in single-cell rna-seq experiments. Genome Biol. 2016;17(1):222.
38. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2.
39. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell rna-sequencing experiments. Biostatistics. 2018;19(4):562–78.
40. Van den Berge K, Perraudeau F, Soneson C, Love MI, Risso D, Vert J-P, Robinson MD, Dudoit S, Clement L. Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. Genome Biol. 2018;19(1):1–17.
41. Deaton AM, Webb S, Kerr ARW, Illingworth RS, Guy J, Andrews R, Bird A. Cell type–specific dna methylation at intragenic cpg islands in the immune system. Genome Res. 2011;21(7):1074–86.
42. Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimr: power analysis for bulk and single cell rna-seq experiments. Bioinformatics. 2017;33(21):3486–8.
43. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. Nat Biotechnol. 2020;38(6):737–46.
44. Van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(3):716–29.
45. Li WV, Li JJ. An accurate and robust imputation method scimpute for single-cell rna-seq data. Nat Commun. 2018;9(1):1–9.

46. Pierson E, Yau C. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015;16(1):1–10.
47. Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. Drimpute: imputing dropout events in single cell rna sequencing data. BMC Bioinformatics. 2018;19(1):1–10.
48. Mo H, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray JI, Raj A, Li M, Zhang NR. Saver: gene expression recovery for single-cell rna sequencing. Nat Methods. 2018;15(7):539–42.
49. Talwar D, Mongia A, Sengupta D, Majumdar A. Autoimpute: Autoencoder based imputation of single-cell rna-seq data. Sci Rep. 2018;8(1):1–11.
50. Ronen J, Akalin A. netsmooth: Network-smoothing based imputation for single cell rna-seq. F1000Research. 2018;7:7.
51. Badsha MdB, Li R, Liu B, Li YI, Xian M, Banovich NE, Fu AQ. Imputation of single-cell gene expression with an autoencoder neural network. Quant Biol. 2020;8(1):78–94.
52. Eraslan G, Simon LM, Mircea M, Mueller NS, Theis FJ. Single-cell rna-seq denoising using a deep count autoencoder. Nat Commun. 2019;10(1):1–14.
53. Mongia A, Sengupta D, Majumdar A. Mcimpute: Matrix completion based imputation for single cell rna-seq data. Front Genet. 2019;10:9.
54. Chen C, Changjing W, Linjie W, Wang X, Deng M, scrmd RX. Imputation for single cell rna-seq data via robust matrix decomposition. Bioinformatics. 2020;36(10):3156–61.
55. Yang MQ, Weissman SM, Yang W, Zhang J, Canaann A, Guan R. Misc: missing imputation for single-cell rna sequencing data. BMC Syst Biol. 2018;12(7):114.
56. Tang W, Bertaux F, Thomas P, Stefanelli C, Saint M, Marguerat S, Shahrezaei V. baynorm: Bayesian gene expression recovery, imputation and normalization for single-cell rna-sequencing data. Bioinformatics. 2020;36(4):1174–81.
57. Elyanow R, Dumitrascu B, Engelhardt BE, Raphael BJ. netnmf-sc: leveraging gene–gene interactions for imputation and dimensionality reduction in single-cell expression analysis. Genome Res. 2020;30(2):195–204.
58. Moussa M, Măndoiu II. Locality sensitive imputation for single cell rna-seq data. J Comput Biol. 2019;26(8):822–35.
59. Peng T, Zhu Q, Yin P, Tan K. Scrabble: single-cell rna-seq imputation constrained by bulk rna-seq data. Genome Biol. 2019;20(1):88.
60. Xu Y, Zhang Z, You L, Liu J, Fan Z, Zhou X. scigans: single-cell rna-seq imputation using generative adversarial networks. Nucleic Acids Res. 2020;48(15):e85—e85.
61. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8.
62. Arisdakessian C, Poirion O, Yunits B, Zhu X, Garmire LX. Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. Genome Biol. 2019;20(1):1–14.
63. Chen M, Zhou X. Viper: variability-preserving imputation for accurate gene expression recovery in single-cell rna sequencing studies. Genome Biol. 2018;19(1):1–15.
64. Svensson V. Droplet scrna-seq is not zero-inflated. Nat Biotechnol. 2020;38(2):147–50.
65. Kim TH, Zhou X, Chen M. Demystifying “drop-outs” in single-cell umi data. Genome Biol. 2020;21(1):1–19.
66. Qiu P. Embracing the dropouts in single-cell rna-seq analysis. Nat Commun. 2020;11(1):1–9.
67. Silverman JD, Roche K, Mukherjee S, David LA. Naught all zeros in sequence count data are the same. Comput Struct Biotechnol J. 2020;18:2789.
68. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell rna sequencing data. Genome Biol. 2017;18(1): 1–15.
69. Alberts B, Johnson A, Lewis J, Morgan D, Raff M Roberts, et al. Molecular biology of the cell. London: Garland Science, Taylor and Francis Group; 2018.
70. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mrna synthesis in mammalian cells. PLoS Biol. 2006;4(10):e309.
71. Sanchez A, Golding I. Genetic determinants and cellular constraints in noisy gene expression. Science. 2013;342(6163):1188–93.
72. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. Mammalian genes are transcribed with widely different bursting kinetics. Science. 2011;332(6028):472–4.
73. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13(9):613–26.
74. Inukai S, Kock KH, Bulyk ML. Transcription factor–dna binding: beyond binding site motifs. Curr Opin Genet Dev. 2017;43:110–9.
75. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. The human transcription factors. Cell. 2018;172(4):650–65.
76. Paszek P. Modeling stochasticity in gene regulation: characterization in the terms of the underlying distribution function. Bull Math Biol. 2007;69(5):1567–601.
77. Peccoud J, Ycart B. Markovian modeling of gene-product synthesis. Theor Popul Biol. 1995;48(2):222–34.
78. Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell rna-sequencing data. Genome Biol. 2013;14(1):1–12.
79. Schwaber J, Andersen S, Nielsen L. Shedding light: the importance of reverse transcription efficiency standards in data interpretation. Biomol Detect Quantif. 2019;17:100077.
80. Bustin S, Dhillon HS, Kirvell S, Greenwood C, Parker M, Shipley GL, Nolan T. Variability of the reverse transcription step: practical implications. Clin Chem. 2015;61(1):202–12.
81. Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. Front Microbiol. 2017;8:2114.
82. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. Primer-directed enzymatic amplification of dna with a thermostable dna polymerase. Science. 1988;239(4839):487–91.
83. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, Zettel M, Coleman P. Analysis of gene expression in single live neurons. Proc Natl Acad Sci. 1992;89(7):3010–4.

84. Tang F, Lao K, Surani MA. Development and applications of single-cell transcriptome analysis. Nat Methods. 2011;8(4):S6—S11.
85. Fu Y, Wu P-H, Beane T, Zamore PD, Weng Z. Elimination of pcr duplicates in rna-seq and small rna-seq using unique molecular identifiers. BMC Genom. 2018;19(1):531.
86. Tung P-Y, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, Gilad Y. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7:39921.
87. Shiroguchi K, Jia TZ, Sims PA, Xie XS. Digital rna sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci. 2012;109(4):1347–52.
88. Cha RS, Thilly WG. Specificity, efficiency, and fidelity of pcr. PCR Methods Appl. 1993;3(3):18–29.
89. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. Nucleic Acids Res. 2008;36(16):e105.
90. Smith T, Heger A, Sudbery I. Umi-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome Res. 2017;27(3):491–9.
91. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. Genome Biol. 2011;12(2):1–14.
92. Dueck HR, Ai R, Camarena A, Ding B, Dominguez R, Evgrafov OV, Fan J-B, Fisher SA, Herstein JS, Kim TK, et al. Assessing characteristics of rna amplification methods for single cell rna sequencing. BMC Genom. 2016;17(1):1–22.
93. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. Genome Biol. 2019;20(1):1–16.
94. Sarkar A, Stephens M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. Nat Genet. 2021;53(6):770–7.
95. Zhu L, Lei J, Devlin B, Roeder K. A unified statistical framework for single cell and bulk rna sequencing data. Ann Appl Stat. 2018;12(1):609.
96. Zand M, Ruan J. Network-based single-cell rna-seq data imputation enhances cell type identification. Genes. 2020;11(4):377.
97. Di R, Zhang S, Lytal N, An L. scdoc: correcting drop-out events in single-cell rna-seq data. Bioinformatics. 2020;36(15):4233–9.
98. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, et al. Eleven grand challenges in single-cell data science. Genome Biol. 2020;21(1):1–35.
99. Andrews TS, Hemberg M. False signals induced by single-cell imputation. F1000Research. 2018;7:7.
100. 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. Nature Commun. 2018;9(1):1–17.
101. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, Lönnerberg P, Linnarsson S. Quantitative single-cell rna-seq with unique molecular identifiers. Nat Methods. 2014;11(2):163.
102. Sun T, Song D, Li WV, Li JJ. scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. Genome Biol. 2021;22(1):1–37.
103. Zhang X, Chenling X, Yosef N. Simulating multiple faceted variability in single cell rna sequencing. Nat Commun. 2019b;10(1):1–16.
104. Dibaeinia P, Sinha S. Sergio: a single-cell expression simulator guided by gene regulatory networks. Cell Syst. 2020;11(3):252–71.
105. Ge X, Chen YE, Song D, McDermott M, Woyshner K, Manousopoulou A, Wang N, Li W, Wang LD, Li JJ. Clipper: p-value-free FDR control on high-throughput data from two conditions. Genome Biol. 2021;22(1):1–29.
106. Amodio M, Van Dijk D, Srinivasan K, Chen WS, Mohsen H, Moon KR, Campbell A, Zhao Y, Wang X, Venkataswamy M, et al. Exploring single-cell data with deep multitasking neural networks. Nat Methods. 2019;16(11):1139–45.
107. Warton DI. Why you cannot transform your way out of trouble for small counts. Biometrics. 2018;74(1):362–8.
108. Andrew. You should (usually) log transform your positive data. 2019. https://statmodeling.stat.columbia.edu/2019/08/21/you-should-usually-log-transform-your-positive-data/. Accessed 21 Dec 2020.
109. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. Genome Biol. 2014;15(2):R29.
110. Zhang L, Zhang S. Comparison of computational methods for imputing single-cell rna-sequencing data. IEEE/ACM Trans Comput Biol Bioinforma. 2018;17(2):376–89.
111. He Y, Yuan H, Cheng W, Xie Z. Disc: a highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning. Genome Biol. 2020;21(1):1–28.
112. Li Y, Ge X, Peng F, Li W, Li JJ. A large-sample crisis? exaggerated false positives by popular differential expression methods. bioRxiv. 2021. https://doi.org/10.1101/2021.08.25.457733.
113. Li R, Quon G. scbfa: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. Genome Biol. 2019;20(1):193.
114. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nature Biotechnol. 2015;33(3):269–76.
115. Chen H, Guo J, Mishra SK, Robson P, Niranjan M, Zheng J. Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development. Bioinformatics. 2015;31(7):1060–6.
116. Lim CY, Wang H, Woodhouse S, Piterman N, Wernisch L, Fisher J, Göttgens B. Btr: training asynchronous boolean models using single-cell expression data. BMC Bioinformatics. 2016;17(1):1–18.
117. Agrawal A, Chiu AM, Le M, Halperin E, Sankararaman S. Scalable probabilistic pca for large-scale genetic variation data. PLoS Genetics. 2020;16(5):e1008773.
118. Hou W, Ji Z, Ji H, Hicks SC. A systematic evaluation of single-cell rna-sequencing imputation methods. bioRxiv. 2020;21(1):1–30.
119. Baker SC, Bauer SR, Beyer RP, Brenton JD, Bromley B, Burrill J, Causton H, Conley MP, Elespuru R, Fero M, et al. The external rna controls consortium: a progress report. Nat Methods. 2005;2(10):731.
120. Chen W, Zhao Y, Chen X, Yang Z, Xu X, Bi Y, Chen V, Li J, Choi H, Ernest B, Tran B, Mehta M, Kumar P, Farmer A, Mir A, Mehra UA, Li JL, Moos Jr M, Xiao W, Wang C. A multicenter study benchmarking single-cell RNA

sequencing technologies using reference samples. Nat Biotechnol. 2021;39(9):1103–14. https://doi.org/10.1038/s41587-020-00748-9.

121. Andrews TS, Kiselev VY, McCarthy D, Hemberg M. Tutorial: guidelines for the computational analysis of single-cell rna sequencing data. Nat Protoc. 2021;16(1):1–9.

122. Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. 2022. https://doi.org/10.5281/zenodo.4393040.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.