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,*]

Abstract

Single-cell RNA sequencing (scRNA-seq) technologies have revolutionized biomedical sciences by enabling genome-wide profiling of gene expression levels at an unprecedented single-cell resolution. A distinct characteristic of scRNA-seq data is the vast proportion of zeros unseen in bulk RNA-seq data. Researchers view these zeros differently: some regard zeros as biological signals representing no or low gene expression, while others regard zeros as false signals or missing data to be corrected. As a result, the scRNA-seq field faces much controversy regarding how to handle zeros in data analysis. In this paper, we first discuss the sources of biological and non-biological zeros in scRNA-seq data. Second, we evaluate the impacts of non-biological zeros on cell clustering and differential gene expression analysis. Third, we summarize the advantages, disadvantages, and suitable uses of three input data types: observed counts, imputed counts, and binarized counts. Fourth, we benchmark the performance of the three input data types in cell clustering, dimension reduction, and differential gene expression analysis. Finally, we discuss the open questions regarding non-biological zeros and 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

[1] Department of Statistics, University of California, Los Angeles, CA 90095-1554
[2] Bioinformatics Interdepartmental Ph.D. Program, University of California, Los Angeles, CA 90095-7246
[3] Department of Human Genetics, University of California, Los Angeles, CA 90095-7088
[4] Department of Computational Medicine, University of California, Los Angeles, CA 90095-1766
[*] To whom correspondence should be addressed. Email: jli@stat.ucla.edu

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, 47–62]. More recently, however, there were voices against the use of zero-inflated models for scRNA-seq data generated by UMI protocols [63]. Besides, there was a proposal for treating zeros as useful information that researchers should embrace [64]. 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.

Table1 summarizes the key concepts used in this paper, including their definitions and categories (biology, technology, and modeling).

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,

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 (synthesised 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
techinical zero	absence of reads or UMIs of a gene in a cell due to the library-preparation steps (e.g. cDNA amplification) before sequencing	sequencing technology
sampling zero	absence of reads or UMIs of a gene in a cell due to 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 1: A summary of the key concepts used in this paper, including their definitions and nature.

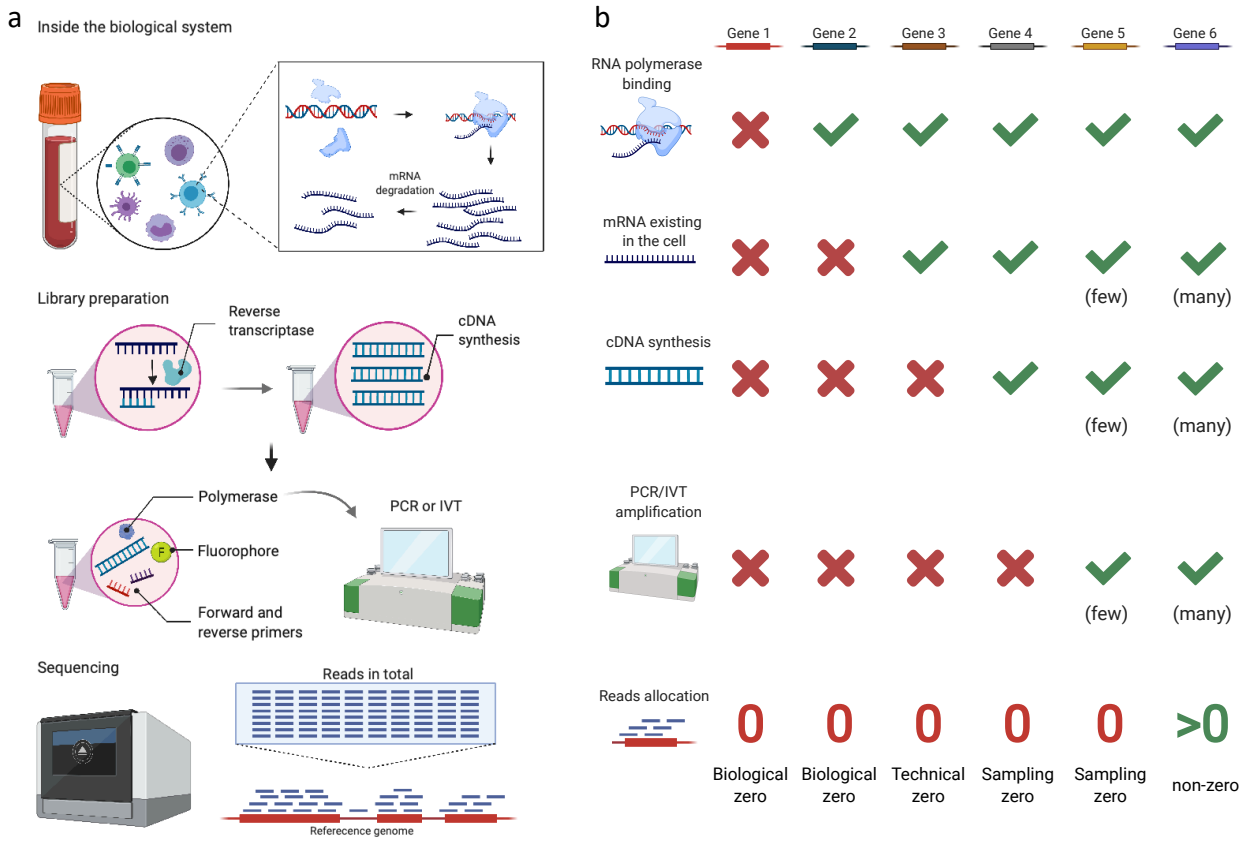


Figure 1: **Sources of zeros in scRNA-seq data. (a)** An overview of a scRNA-seq experiment. Biological factors that determine true gene expression levels include transcription and mRNA degradation (top panel). Technical procedures that affect gene expression measurements include cDNA synthesis, PCR 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 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 `BioRender.com`.

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. [65] except a slight difference (we refer to the zeros due to inefficient amplificaiton, 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. Fig. 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. Fig. 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 and non-biological zeros are hardly distinguishable in scRNA-seq data without biological knowledge or spike-in control (see Future directions).

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 [65]. 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 [66, 67]. 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, 68–70]. Specifically, in eukaryotic cells, transcription is initiated by the binding of specific transcription factors (TFs) and RNA polymerase to the promoter of a gene [71–73]. Due to the stochasticity of TF binding, a gene switches between active and inactive states, and its transcription only occurs during the active state [74]. 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 [74–76]. Fig. 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

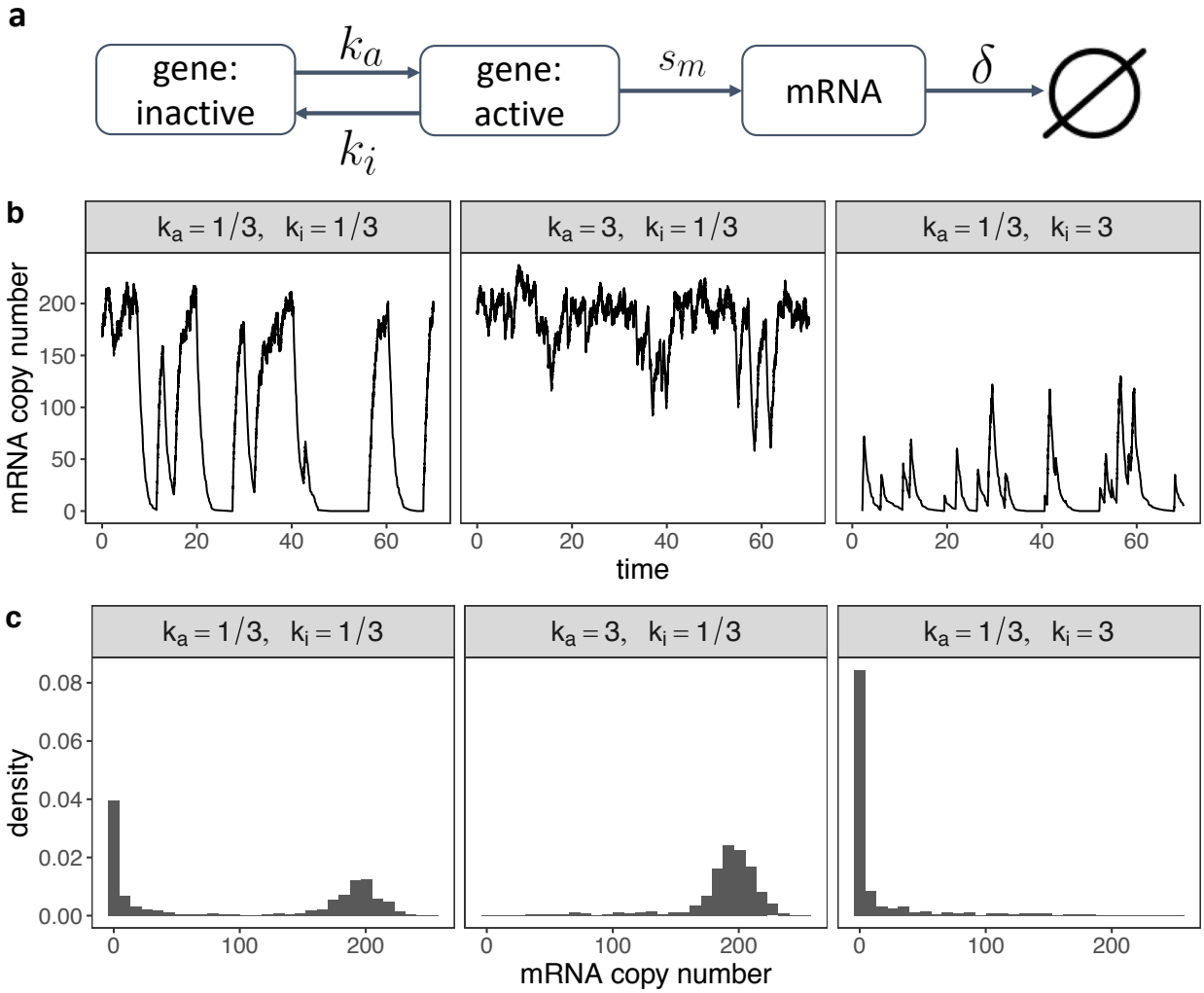


Figure 2: **A two-state stochastic model of the expression levels of one gene. (a)** A diagram of the two-state gene expression model [74–76], where a gene stochastically switches from an inactive state to an active state at rate k_a and from an active state to an inactive state at rate k_i. The gene 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.

in a cell. There are two types of non-biological zeros [65]: (1) technical zeros, which arise from library-preparation steps before sequencing, and (2) sampling zeros, which result from 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% [77], depending on multiple experimental parameters [78]. 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 [63, 79], 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 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) [80] 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 [81]. However, compared with PCR, IVT requires more input cDNAs to ensure successful amplification; thus, PCR is still the dominant amplification procedure for scRNA-seq [82]. Though indispensable, cDNA amplification is known to introduce biases into cDNA copy numbers because the amplification efficiency depends on cDNA sequence and structure [83, 84]. For example, GC-rich cDNA sequences are harder to be amplified [65]. 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 [85, 86]. Moreover, cDNA copy number biases would accumulate as the number of amplification cycles increases [83, 87]; that is, the more cycles, the larger the difference of two genes (with different amplification efficiency) in cDNA copy numbers [88, 89]. Since different scRNA-seq experiments may use different numbers of amplification cycles (e.g., 18 cycles in a Smart-seq2 experiment [90] and 14 cycles in a 10x Genomics experiment [91]), cDNA copy number biases differ among scRNA-seq datasets. In addition, the non-linear amplification nature

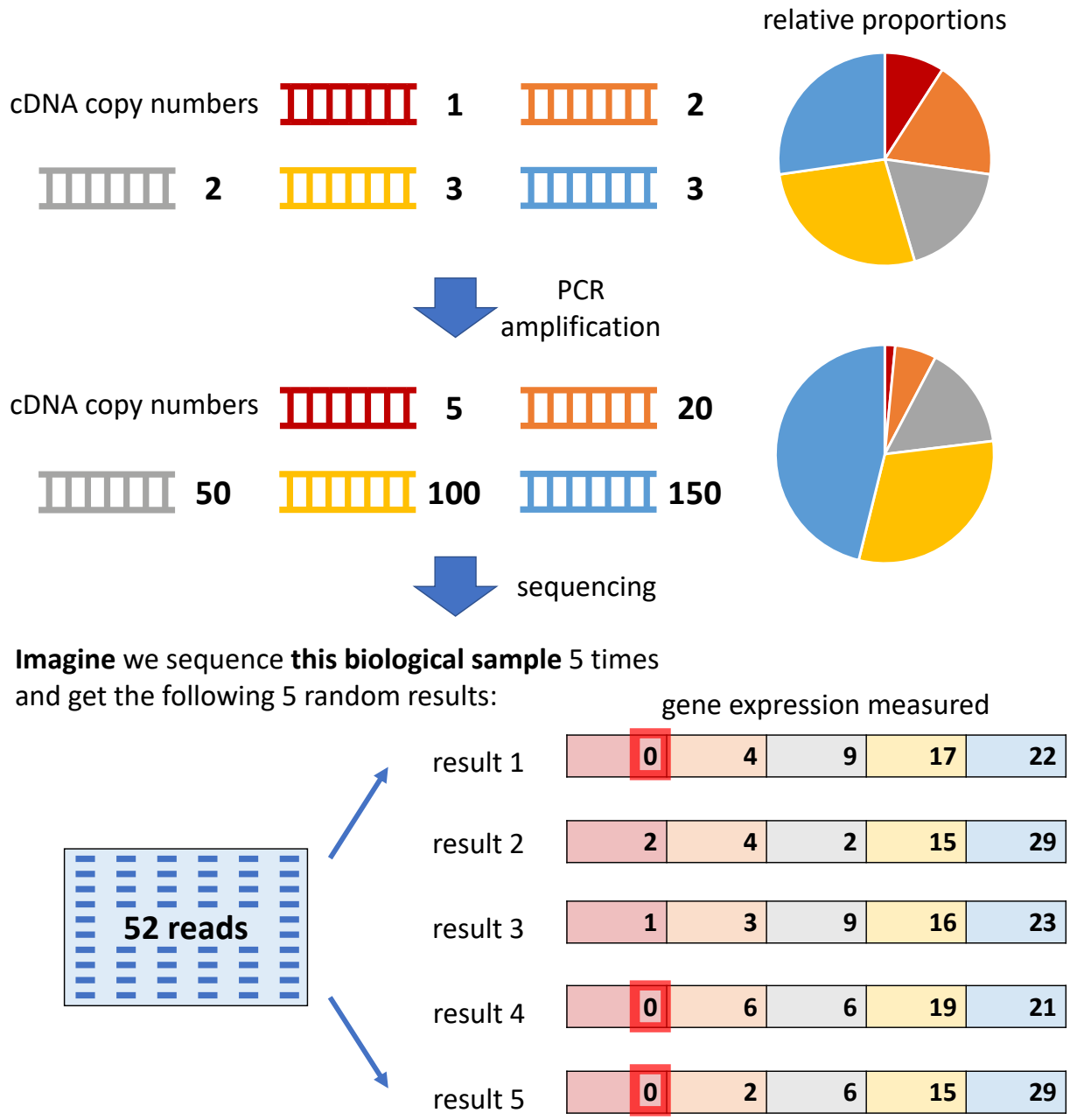


Figure 3: **A toy example showing how the PCR amplification may result in sampling zeros.** Five genes have their cDNAs amplified by PCR. After the non-linear amplification, their relative proportions change. If the sequencing depth is limited to 52 reads, the first gene has sampling zeros in three out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

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).

Box1: 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 [92]. 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, 93, 94]; some meant non-biological zeros and low expression measurements [45, 95]; some meant all zeros [46, 47, 96]. In addition, "dropout" was often used as an adjective to mean the existence of many zeros [97]. Such inconsistent uses of "dropouts" are emphasized in a recent work [92]. 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, 94]; some meant the additional zeros that cannot be explained by the negative binomial (NB) model [95]. 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 [93]. 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.

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, 98]. 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, 91, 92].

The use of UMIs in scRNA-seq can correct the amplification biases in non-zero gene expression measurements [99]; 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 [98]. 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 one (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 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 using droplet-based ERCC spike-in data. To evaluate this claim on real scRNA-seq data

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)	$\lambda \cdot (1-p)$	$\lambda \cdot (1-p) + \lambda^2 \cdot (1-p) \cdot (p + \frac{1}{\psi})$	$(\frac{\psi}{\psi+\lambda})^{\psi} \cdot (1-p) + p$

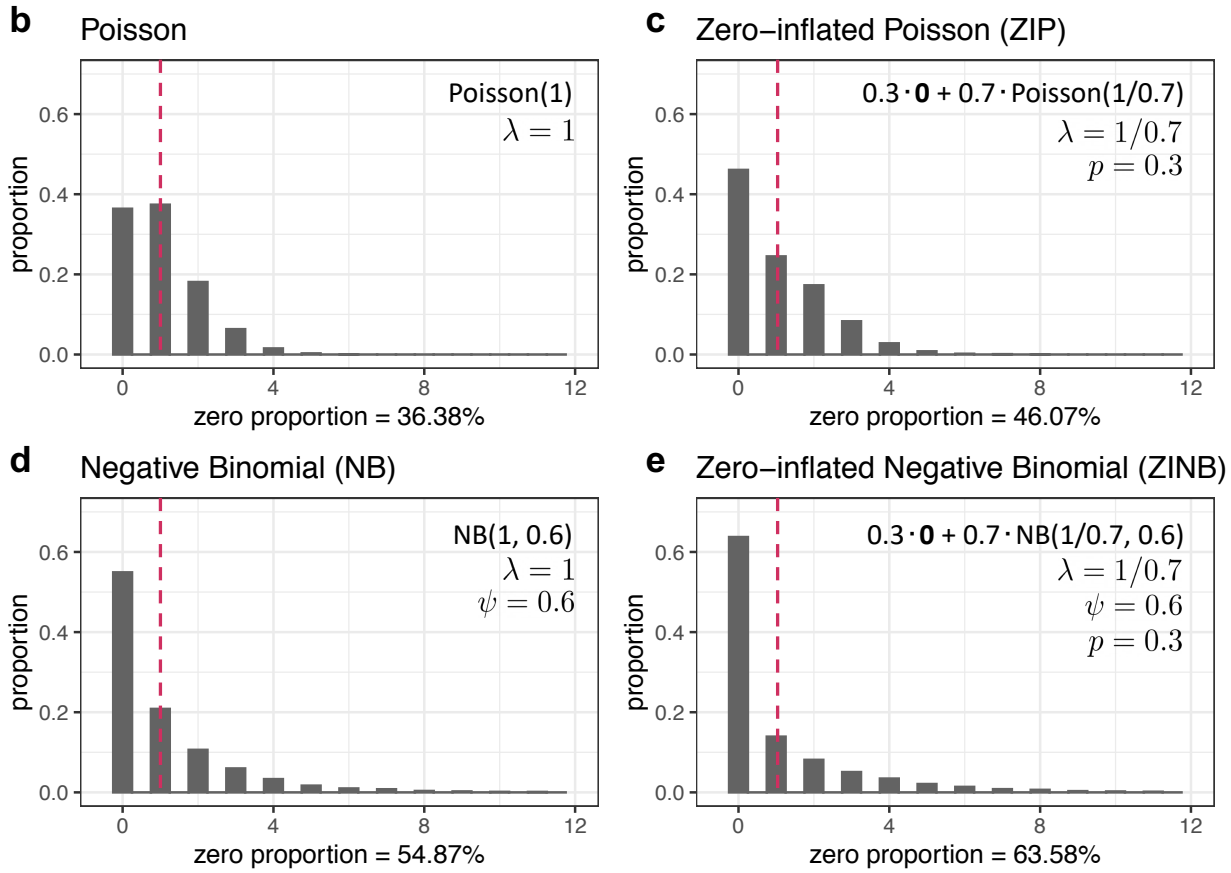


Figure 4: **Four count distributions: Poisson, zero-inflated Poisson (ZIP), negative binomial (NB), and zero-inflated negative binomial (ZINB). (a)** Parameterization, mean, variance, and zero proportion of each of the four distributions. **(b)**, **(c)**, **(d)**, and **(e)** Illustration of the probability 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.

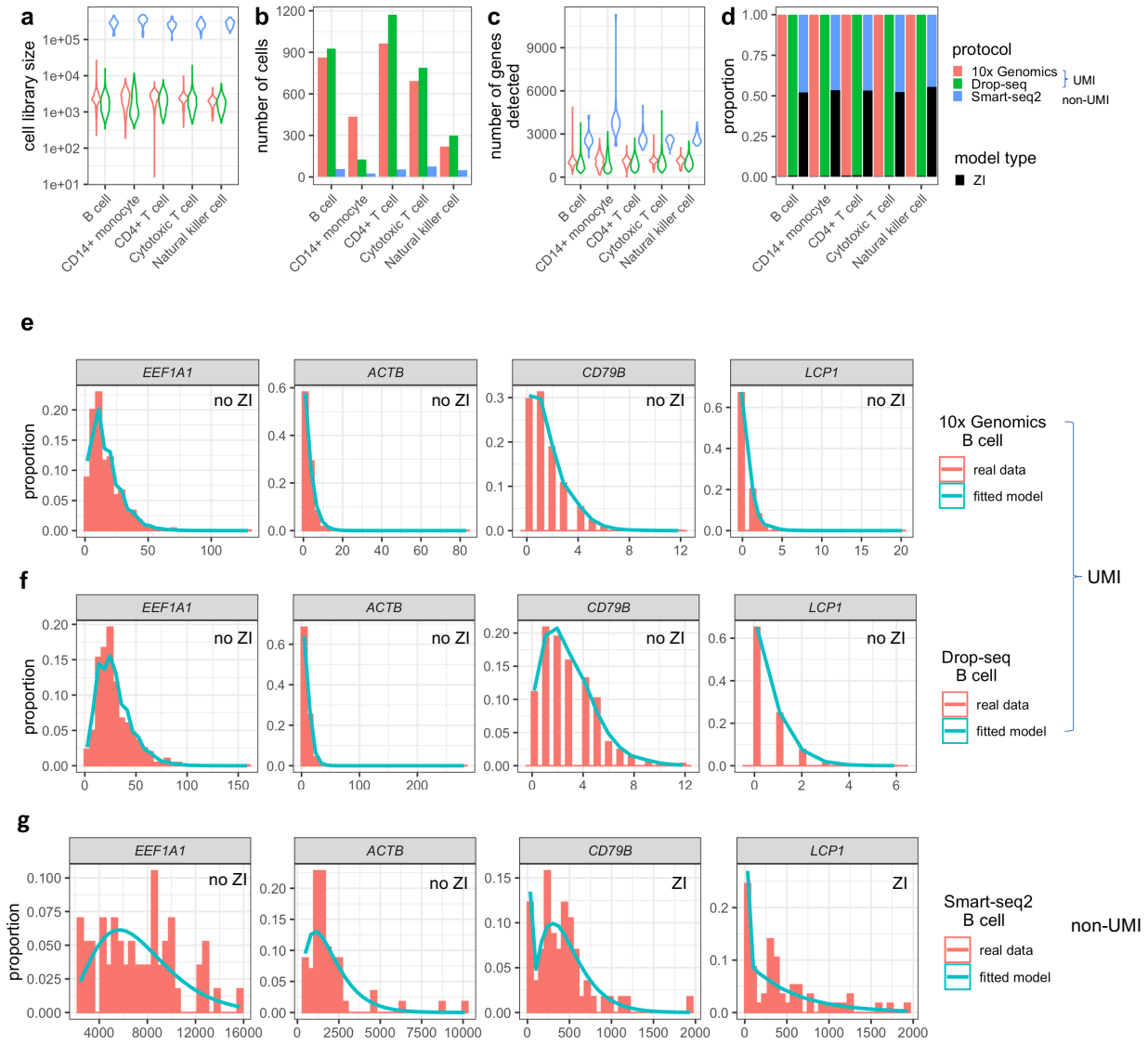


Figure 5: **Statistical modeling of 10x Genomics, Drop-seq, and Smart-seq2 data for the same PMBC sample.** 10x Genomics and Drop-seq data are UMI-based, while Smart-seq2 data are non-UMI-based. **(a)** Violin plots showing the distribution of cell library sizes for each of five PMBC cell types measured 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' distributions of UMI counts 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' distributions of UMI counts in B cells measured by Drop-seq. **(g)** The same four example genes' distributions of read counts in B cells measured by Smart-seq2. ZI models are chosen for two genes.

with 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 to be the same across protocols. 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. Fig. 5a–c show that, compared with the two UMI-based datasets, the Smart-seq2 (non-UMI-based) dataset has larger cell library sizes, 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 [100] for detail). The rationale is to choose the least complex distribution that fits the data well. Fig. 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 [63, 92], 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

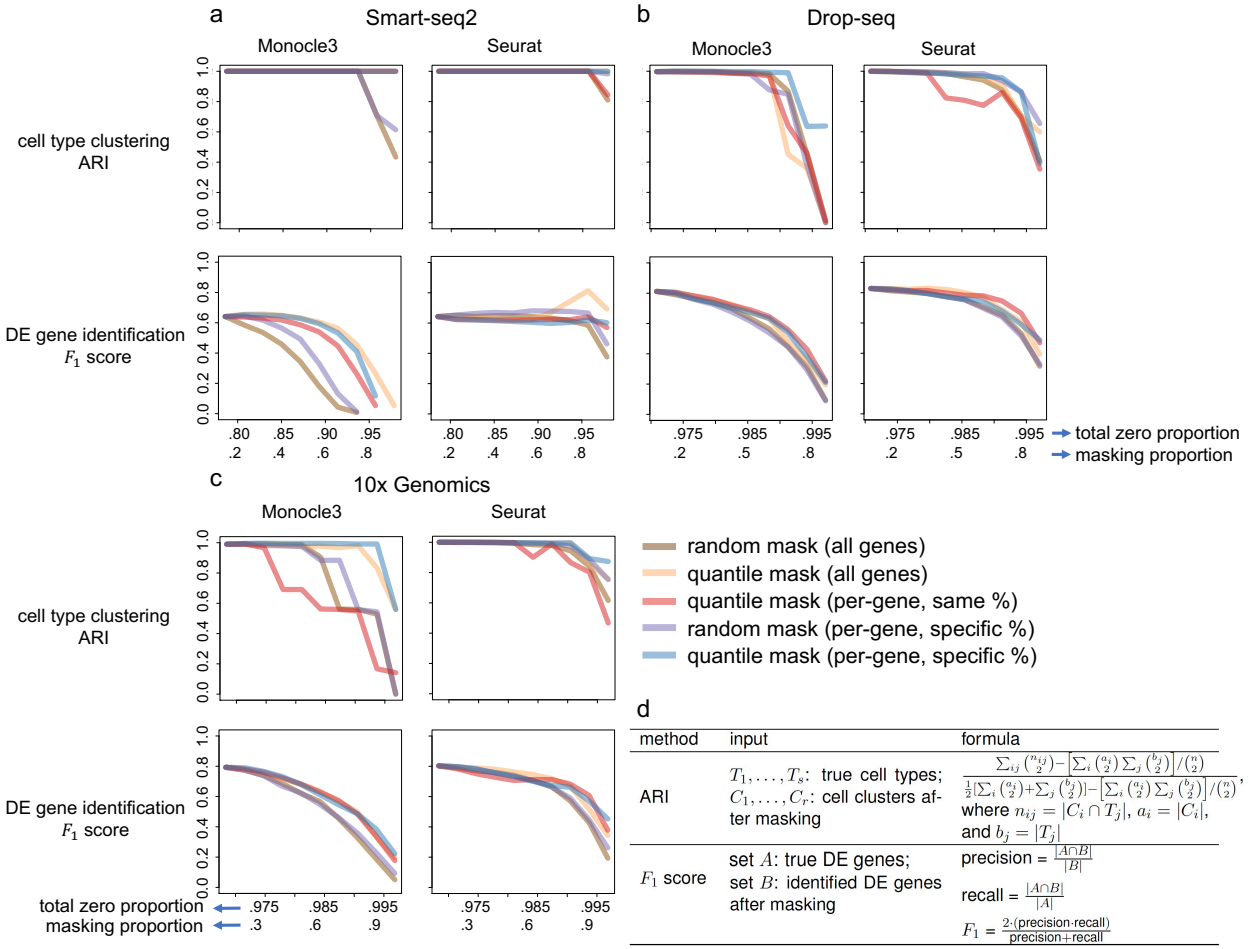
To evaluate the effects of non-biological zeros on scRNA-seq data analysis, such as cell clustering and DE gene identification, we need access to true cell types and true DE genes. Hence, we use scDesign2 [100], a probabilistic, flexible simulator we developed to generate realistic scRNA-seq count data 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).

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 Supplementary), 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) $\times 5$ (# 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. Fig. 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 %)` (Fig. S1b)—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 [64], an assumption aligned with these two masking schemes. Finally, among the three protocols, clustering on Smart-seq2 data is



method	input	formula
ARI	$T_1, \ldots, T_s$: true cell types; $C_1, \ldots, C_r$: cell clusters after masking	$\frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]/\binom{n}{2}}$, 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 = $\frac{2 \cdot (\text{precision} \cdot \text{recall})}{\text{precision} + \text{recall}}$

Figure 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.

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 (all genes)` and `random mask (per-gene, specific %)` (Fig. S1b)—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 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 Supplementary Fig. S2. It is worth noting that although the false discovery rate is set to 5%, the precision rates of both Monocle3 and Seurat are 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 [101]. In addition, compared to Seurat, Monocle3 shows a greater fluctuation in both precision and recall as the masking proportion increases.

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, 102]. 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

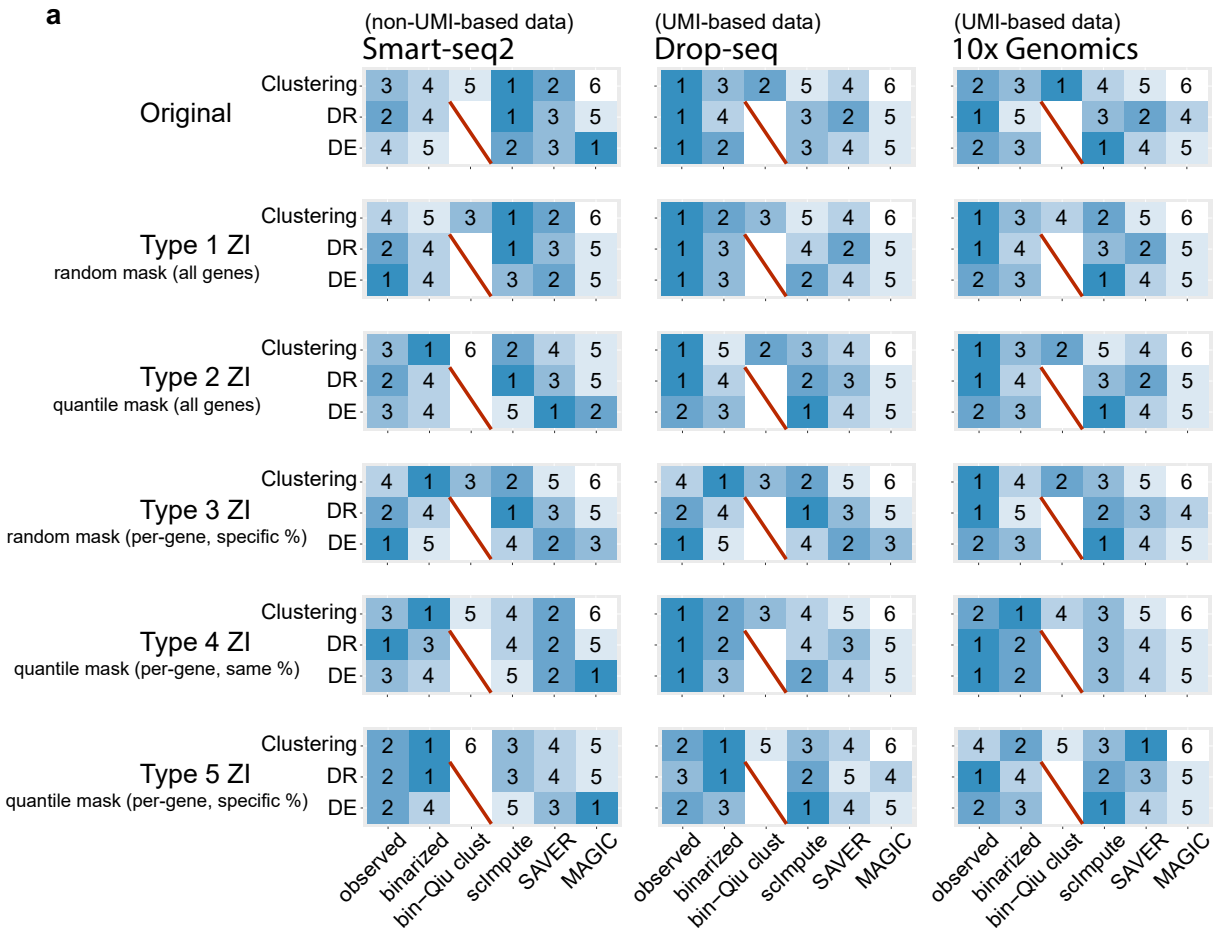
Current scRNA-seq data analysis typically takes three types of input data: observed, imputed, and binarized counts. Researchers use imputed and binarized counts to deal with the vast proportion of zeros. Although log-transformed counts are often used as input data, this practice is under controversy [91, 103, 104] 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, 105]. 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, 47–62]. Of course, imputation has the risk of biasing data, leading to false signals [97] or diminished biological variation [45, 106]. 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, 107]. 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, 108], conditional on verified false discovery rate control [101, 109].

Moreover, a recent article provides a new perspective by proposing to use only binarized counts (with all non-zero counts truncated as ones) for cell clustering [64]. 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 mean-

a



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

Figure 7: (Continued on the following page.)

Figure 7: **Performance ranks of using observed, binarized, and imputed counts (from three experimental protocols ad 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 4–6 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 7–9 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.

ingful 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, 110]. 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 [111–113]. 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 [64]. 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 [114].

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 benchmark study [115].

To benchmark cell clustering and dimension reduction results, we use the three real scRNA-seq PBMC datasets with labelled 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 [100] from two cell types (CD4+ T cells and cytotoxic T cells) in the three real datasets.

For cell clustering, we use two algorithms: Qiu's algorithm designed specifically for binarized counts [64] 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 (Fig. S3; top row).

For cell dimension reduction, we perform UMAP (implemented in Seurat) on the observed, binarized and imputed counts (Figs. S4, S5, and S6 show the results of Smart-seq2, Drop-seq, and 10x Genomics, respectively). We use the average Silhouette score to evaluate how well the labeled cell types are separated in the two-dimensional UMAP space (Fig. S7; top row).

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 (Fig. S9), recall (Fig. S10) and F_1 score (Fig. S8) to evaluate the identification results.

Fig. 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 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 Supplementary) to introducing additional non-biological zeros. Under each masking scheme, we mask 50% of the original non-zero 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. 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 non-biological zeros introduced.

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 [116]), 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 [117]. 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 [118].

Code availability

The R code for reproducing the results in Figs. 2–6 is available at `https://doi.org/10.5281/zenodo.4393041`.

Funding

This work was supported by the following grants: National Science Foundation DBI-1846216, NIH/NIGMS R01GM120507, 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.).

Conflicts of interests

The authors declare no conflicts of interests.

Acknowledgements

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

References

[1] Antoine-Emmanuel Saliba, Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Single-cell rna-seq: advances and future challenges. Nucleic acids research, 42(14):8845–8860, 2014.

[2] Serena Liu and Cole Trapnell. Single-cell transcriptome sequencing: recent advances and remaining challenges. F1000Research, 5, 2016.

[3] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. Nature Reviews Genetics, 20(5):273–282, 2019.

[4] Sophie Tritschler, Maren Büttner, David S Fischer, Marius Lange, Volker Bergen, Heiko Lickert, and Fabian J Theis. Concepts and limitations for learning developmental trajectories from single cell genomics. Development, 146(12):dev170506, 2019.

[5] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly

parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161(5):1202–1214, 2015.

[6] Robert Salomon, Dominik Kaczorowski, Fatima Valdes-Mora, Robert E Nordon, Adrian Neild, Nona Farbehi, Nenad Bartonicek, and David Gallego-Ortega. Droplet-based single cell rnaseq tools: a practical guide. Lab on a Chip, 19(10):1706–1727, 2019.

[7] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. Nature communications, 8(1):1–12, 2017.

[8] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. Nature protocols, 9(1):171–181, 2014.

[9] Alex A Pollen, Tomasz J Nowakowski, Joe Shuga, Xiaohui Wang, Anne A Leyrat, Jan H Lui, Nianzhen Li, Lukasz Szpankowski, Brian Fowler, Peilin Chen, et al. Low-coverage single-cell mrna sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature biotechnology, 32(10):1053, 2014.

[10] Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. Power analysis of single-cell rna-sequencing experiments. Nature methods, 14(4):381–387, 2017.

[11] Xiannian Zhang, Tianqi Li, Feng Liu, Yaqi Chen, Jiacheng Yao, Zeyao Li, Yanyi Huang, and Jianbin Wang. Comparative analysis of droplet-based ultra-high-throughput single-cell rna-seq systems. Molecular cell, 73(1):130–142, 2019.

[12] Fang Wang, Shaoheng Liang, Tapsi Kumar, Nicholas Navin, and Ken Chen. Scmarker: ab initio marker selection for single cell transcriptome profiling. PLoS computational biology, 15(10):e1007445, 2019.

[13] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. Nature biotechnology, 33(5):495–502, 2015.

[14] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. Sc3: consensus clustering of single-cell rna-seq data. Nature methods, 14(5):483–486, 2017.

[15] Minzhe Guo, Hui Wang, S Steven Potter, Jeffrey A Whitsett, and Yan Xu. Sincera: a pipeline for single-cell rna-seq profiling analysis. PLoS computational biology, 11(11):e1004575, 2015.

[16] Yu-Jui Ho, Naishitha Anaparthy, David Molik, Grinu Mathew, Toby Aicher, Ami Patel, James Hicks, and Molly Gale Hammell. Single-cell rna-seq analysis identifies markers of resistance to targeted braf inhibitors in melanoma cell populations. Genome research, 28(9):1353–1363, 2018.

[17] Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. Science, 347(6226):1138–1142, 2015.

[18] Roser Vento-Tormo, Mirjana Efremova, Rachel A Botting, Margherita Y Turco, Miquel Vento-Tormo, Kerstin B Meyer, Jong-Eun Park, Emily Stephenson, Krzysztof Polański, Angela Goncalves, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. Nature, 563(7731):347–353, 2018.

[19] Adam P Croft, Joana Campos, Kathrin Jansen, Jason D Turner, Jennifer Marshall, Moustafa Attar, Loriane Savary, Corinna Wehmeyer, Amy J Naylor, Samuel Kemble, et al. Distinct fibroblast subsets drive inflammation and damage in arthritis. Nature, 570(7760):246–251, 2019.

[20] Peijie Lin, Michael Troup, and Joshua WK Ho. Cidr: Ultrafast and accurate clustering through imputation for single-cell rna-seq data. Genome biology, 18(1):59, 2017.

[21] Zhe Sun, Ting Wang, Ke Deng, Xiao-Feng Wang, Robert Lafyatis, Ying Ding, Ming Hu, and Wei Chen. Dimm-sc: a dirichlet mixture model for clustering droplet-based single cell transcriptomic data. Bioinformatics, 34(1):139–146, 2018.

[22] Christopher Yau et al. pcareduce: hierarchical clustering of single cell transcriptional profiles. BMC bioinformatics, 17(1):140, 2016.

[23] Tallulah S Andrews and Martin Hemberg. M3drop: dropout-based feature selection for scrnaseq. Bioinformatics, 35(16):2865–2867, 2019.

[24] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology, 32(4):381, 2014.

[25] Zhicheng Ji and Hongkai Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. Nucleic acids research, 44(13):e117–e117, 2016.

[26] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC genomics, 19(1):477, 2018.

[27] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah A Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell trajectories. Nature methods, 14(10):979, 2017.

[28] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature, 566(7745):496–502, 2019.

[29] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. Nature biotechnology, 37(5):547–554, 2019.

[30] Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. Nature methods, 15(4):255, 2018.

[31] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1): 139–140, 2010.

[32] Trung Nghia Vu, Quin F Wills, Krishna R Kalari, Nifang Niu, Liewei Wang, Mattias Rantalainen, and Yudi Pawitan. Beta-poisson model for single-cell rna-seq data analyses. Bioinformatics, 32(14):2128–2135, 2016.

[33] Zhun Miao, Ke Deng, Xiaowo Wang, and Xuegong Zhang. Desingle for detecting three types of differential expression in single-cell rna-seq data. Bioinformatics, 34(18):3223–3224, 2018.

[34] Tomi Suomi, Fatemeh Seyednasrollah, Maria K Jaakkola, Thomas Faux, and Laura L Elo. Rots: An r package for reproducibility-optimized statistical testing. PLoS computational biology, 13(5):e1005562, 2017.

[35] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome biology, 15(12):550, 2014.

[36] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome biology, 16(1):1–13, 2015.

[37] Keegan D Korthauer, Li-Fang Chu, Michael A Newton, Yuan Li, James Thomson, Ron Stewart, and Christina Kendziorski. A statistical approach for identifying differential distributions in single-cell rna-seq experiments. Genome biology, 17(1):222, 2016.

[38] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. Nature methods, 11(7):740–742, 2014.

[39] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. Biostatistics, 19(4):562–578, 2018.

[40] Koen Van den Berge, Fanny Perraudeau, Charlotte Soneson, Michael I Love, Davide Risso, Jean-Philippe Vert, Mark D Robinson, Sandrine Dudoit, and Lieven Clement. Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. Genome biology, 19(1):1–17, 2018.

[41] Aimée M Deaton, Shaun Webb, Alastair RW Kerr, Robert S Illingworth, Jacky Guy, Robert Andrews, and Adrian Bird. Cell type–specific dna methylation at intragenic cpg islands in the immune system. Genome research, 21(7):1074–1086, 2011.

[42] Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsimr: power analysis for bulk and single cell rna-seq experiments. Bioinformatics, 33 (21):3486–3488, 2017.

[43] Jiarui Ding, Xian Adiconis, Sean K Simmons, Monika S Kowalczyk, Cynthia C Hession, Nemanja D Marjanovic, Travis K Hughes, Marc H Wadsworth, Tyler Burks, Lan T Nguyen, et al. Systematic comparison of single-cell and single-nucleus rna-sequencing methods. Nature biotechnology, pages 1–10, 2020.

[44] David van Dijk, Juozas Nainys, Roshan Sharma, Pooja Kaithail, Ambrose J Carr, Kevin R Moon, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data. BioRxiv, page 111591, 2017.

[45] Wei Vivian Li and Jingyi Jessica Li. An accurate and robust imputation method scimpute for single-cell rna-seq data. Nature communications, 9(1):1–9, 2018.

[46] Emma Pierson and Christopher Yau. Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome biology, 16(1):1–10, 2015.

[47] Wuming Gong, Il-Youp Kwak, Pruthvi Pota, Naoko Koyano-Nakagawa, and Daniel J Garry. Drimpute: imputing dropout events in single cell rna sequencing data. BMC bioinformatics, 19(1):1–10, 2018.

[48] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. Nature methods, 15(7):539–542, 2018.

[49] Divyanshu Talwar, Aanchal Mongia, Debarka Sengupta, and Angshul Majumdar. Autoimpute: Autoencoder based imputation of single-cell rna-seq data. Scientific reports, 8(1): 1–11, 2018.

[50] Jonathan Ronen and Altuna Akalin. netsmooth: Network-smoothing based imputation for single cell rna-seq. F1000Research, 7, 2018.

[51] Md Bahadur Badsha, Rui Li, Boxiang Liu, Yang I Li, Min Xian, Nicholas E Banovich, and Audrey Qiuyan Fu. Imputation of single-cell gene expression with an autoencoder neural network. Quantitative Biology, pages 1–17, 2020.

[52] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. Single-cell rna-seq denoising using a deep count autoencoder. Nature communications, 10 (1):1–14, 2019.

[53] Aanchal Mongia, Debarka Sengupta, and Angshul Majumdar. Mcimpute: Matrix completion based imputation for single cell rna-seq data. Frontiers in genetics, 10:9, 2019.

[54] Chong Chen, Changjing Wu, Linjie Wu, Xiaochen Wang, Minghua Deng, and Ruibin Xi. scrmd: Imputation for single cell rna-seq data via robust matrix decomposition. Bioinformatics, 36(10):3156–3161, 2020.

[55] Mary Qu Yang, Sherman M Weissman, William Yang, Jialing Zhang, Allon Canaann, and Renchu Guan. Misc: missing imputation for single-cell rna sequencing data. BMC systems biology, 12(7):114, 2018.

[56] Wenhao Tang, François Bertaux, Philipp Thomas, Claire Stefanelli, Malika Saint, Samuel Marguerat, and Vahid Shahrezaei. baynorm: Bayesian gene expression recovery, imputation and normalization for single-cell rna-sequencing data. Bioinformatics, 36(4):1174–1181, 2020.

[57] Rebecca Elyanow, Bianca Dumitrascu, Barbara E Engelhardt, and Benjamin J Raphael. netnmf-sc: leveraging gene–gene interactions for imputation and dimensionality reduction in single-cell expression analysis. Genome research, 30(2):195–204, 2020.

[58] Marmar Moussa and Ion I Măndoiu. Locality sensitive imputation for single cell rna-seq data. Journal of Computational Biology, 26(8):822–835, 2019.

[59] Tao Peng, Qin Zhu, Penghang Yin, and Kai Tan. Scrabble: single-cell rna-seq imputation constrained by bulk rna-seq data. Genome biology, 20(1):88, 2019.

[60] Yungang Xu, Zhigang Zhang, Lei You, Jiajia Liu, Zhiwei Fan, and Xiaobo Zhou. scigans: single-cell rna-seq imputation using generative adversarial networks. Nucleic acids research, 48(15):e85–e85, 2020.

[61] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. Nature methods, 15(12):1053–1058, 2018.

[62] Cédric Arisdakessian, Olivier Poirion, Breck Yunits, Xun Zhu, and Lana X Garmire. Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. Genome biology, 20(1):1–14, 2019.

[63] Valentine Svensson. Droplet scrna-seq is not zero-inflated. Nature Biotechnology, 38(2): 147–150, 2020.

[64] Peng Qiu. Embracing the dropouts in single-cell rna-seq analysis. Nature communications, 11(1):1–9, 2020.

[65] Justin D Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A David. Naught all zeros in sequence count data are the same. Computational and structural biotechnology journal, 18:2789, 2020.

[66] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. Genome biology, 18(1):1–15, 2017.

[67] Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Peter Walter Keith Roberts, et al. Molecular biology of the cell. Garland Science, Taylor and Francis Group,, 2018.

[68] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mrna synthesis in mammalian cells. PLoS Biol, 4(10):e309, 2006.

[69] Alvaro Sanchez and Ido Golding. Genetic determinants and cellular constraints in noisy gene expression. Science, 342(6163):1188–1193, 2013.

[70] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. science, 332 (6028):472–474, 2011.

[71] François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. Nature reviews genetics, 13(9):613–626, 2012.

[72] Sachi Inukai, Kian Hong Kock, and Martha L Bulyk. Transcription factor–dna binding: beyond binding site motifs. Current opinion in genetics & development, 43:110–119, 2017.

[73] Samuel A Lambert, Arttu Jolma, Laura F Campitelli, Pratyush K Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R Hughes, and Matthew T Weirauch. The human transcription factors. Cell, 172(4):650–665, 2018.

[74] Pawel Paszek. Modeling stochasticity in gene regulation: characterization in the terms of the underlying distribution function. Bulletin of Mathematical Biology, 69(5):1567–1601, 2007.

[75] Jean Peccoud and Bernard Ycart. Markovian modeling of gene-product synthesis. Theoretical population biology, 48(2):222–234, 1995.

[76] Jong Kyoung Kim and John C Marioni. Inferring the kinetics of stochastic gene expression from single-cell rna-sequencing data. Genome biology, 14(1):1–12, 2013.

[77] Jessica Schwaber, Stacey Andersen, and Lars Nielsen. Shedding light: the importance of reverse transcription efficiency standards in data interpretation. Biomolecular detection and quantification, 17:100077, 2019.

[78] Stephen Bustin, Harvinder S Dhillon, Sara Kirvell, Christina Greenwood, Michael Parker, Gregory L Shipley, and Tania Nolan. Variability of the reverse transcription step: practical implications. Clinical Chemistry, 61(1):202–212, 2015.

[79] Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D Peddada. Analysis of microbiome data in the presence of excess zeros. Frontiers in microbiology, 8:2114, 2017.

[80] Randall K Saiki, David H Gelfand, Susanne Stoffel, Stephen J Scharf, Russell Higuchi, Glenn T Horn, Kary B Mullis, and Henry A Erlich. Primer-directed enzymatic amplification of dna with a thermostable dna polymerase. Science, 239(4839):487–491, 1988.

[81] James Eberwine, Hermes Yeh, Kevin Miyashiro, Yanxiang Cao, Suresh Nair, Richard Finnell, Martha Zettel, and Paul Coleman. Analysis of gene expression in single live neurons. Proceedings of the National Academy of Sciences, 89(7):3010–3014, 1992.

[82] Fuchou Tang, Kaiqin Lao, and M Azim Surani. Development and applications of single-cell transcriptome analysis. Nature methods, 8(4):S6–S11, 2011.

[83] Yu Fu, Pei-Hsuan Wu, Timothy Beane, Phillip D Zamore, and Zhiping Weng. Elimination of pcr duplicates in rna-seq and small rna-seq using unique molecular identifiers. Bmc Genomics, 19(1):531, 2018.

[84] Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. Scientific reports, 7:39921, 2017.

[85] Katsuyuki Shiroguchi, Tony Z Jia, Peter A Sims, and X Sunney Xie. Digital rna sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proceedings of the National Academy of Sciences, 109(4):1347–1352, 2012.

[86] Rita S Cha and William G Thilly. Specificity, efficiency, and fidelity of pcr. PCR Methods Appl, 3(3):18–29, 1993.

[87] Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput dna sequencing. Nucleic acids research, 36(16):e105, 2008.

[88] Tom Smith, Andreas Heger, and Ian Sudbery. Umi-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. Genome research, 27(3): 491–499, 2017.

[89] Daniel Aird, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. Genome biology, 12(2): 1–14, 2011.

[90] Hannah R Dueck, Rizi Ai, Adrian Camarena, Bo Ding, Reymundo Dominguez, Oleg V Evgrafov, Jian-Bing Fan, Stephen A Fisher, Jennifer S Herstein, Tae Kyung Kim, et al. Assessing characteristics of rna amplification methods for single cell rna sequencing. BMC genomics, 17(1):1–22, 2016.

[91] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. Genome biology, 20(1):1–16, 2019.

[92] Abhishek K Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single cell rna-seq analysis. BioRxiv, 2020.

[93] Lingxue Zhu, Jing Lei, Bernie Devlin, and Kathryn Roeder. A unified statistical framework for single cell and bulk rna sequencing data. The annals of applied statistics, 12(1):609, 2018.

[94] Maryam Zand and Jianhua Ruan. Network-based single-cell rna-seq data imputation enhances cell type identification. Genes, 11(4):377, 2020.

[95] Di Ran, Shanshan Zhang, Nicholas Lytal, and Lingling An. scdoc: correcting drop-out events in single-cell rna-seq data. Bioinformatics, 36(15):4233–4239, 2020.

[96] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. Genome biology, 21 (1):1–35, 2020.

[97] Tallulah S Andrews and Martin Hemberg. False signals induced by single-cell imputation. F1000Research, 7, 2018.

[98] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. Nature communications, 9(1):1–17, 2018.

[99] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. Nature methods, 11(2):163, 2014.

[100] Tianyi Sun, Dongyuan Song, Wei Vivian Li, and Jingyi Jessica Li. scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. Genome biology, 22(1):1–37, 2021.

[101] Xinzhou Ge, Yiling Elaine Chen, Dongyuan Song, MeiLu McDermott, Kyla Woyshner, Antigoni Manousopoulou, Ning Wang, Wei Li, Leo D Wang, and Jingyi Jessica Li. Clipper: p-value-free fdr control on high-throughput data from two conditions. bioRxiv, pages 2020–11, 2021.

[102] Matthew Amodio, David Van Dijk, Krishnan Srinivasan, William S Chen, Hussein Mohsen, Kevin R Moon, Allison Campbell, Yujiao Zhao, Xiaomei Wang, Manjunatha Venkataswamy, et al. Exploring single-cell data with deep multitasking neural networks. Nature methods, pages 1–7, 2019.

[103] David I Warton. Why you cannot transform your way out of trouble for small counts. Biometrics, 74(1):362–368, 2018.

[104] Andrew. You should (usually) log transform your positive data, August 21st, 2019. URL `https://statmodeling.stat.columbia.edu/2019/08/21/you-should-usually-log-transform-your-positive-data/`.

[105] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. Genome biology, 15(2):R29, 2014.

[106] Lihua Zhang and Shihua Zhang. Comparison of computational methods for imputing single-cell rna-sequencing data. IEEE/ACM transactions on computational biology and bioinformatics, 2018.

[107] Yao He, Hao Yuan, Cheng Wu, and Zhi Xie. Disc: a highly scalable and accurate inference of gene expression and structure for single-cell transcriptomes using semi-supervised deep learning. Genome Biology, 21(1):1–28, 2020.

[108] Kai Simon and Georg Lausen. Viper: augmenting automatic information extraction with visual perceptions. In Proceedings of the 14th ACM international conference on Information and knowledge management, pages 381–388, 2005.

[109] Yumei Li, Xinzhou Ge, Fanglue Peng, Wei Li, and Jingyi Jessica Li. A large-sample crisis? exaggerated false positives by popular differential expression methods. bioRxiv, 2021.

[110] Ruoxin Li and Gerald Quon. scbfa: modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. Genome biology, 20(1):193, 2019.

[111] Victoria Moignard, Steven Woodhouse, Laleh Haghverdi, Andrew J Lilly, Yosuke Tanaka, Adam C Wilkinson, Florian Buettner, Iain C Macaulay, Wajid Jawaid, Evangelia Diamanti,

et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nature biotechnology, 33(3):269–276, 2015.

[112] Haifen Chen, Jing Guo, Shital K Mishra, Paul Robson, Mahesan Niranjan, and Jie Zheng. Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development. Bioinformatics, 31(7):1060–1066, 2015.

[113] Chee Yee Lim, Huange Wang, Steven Woodhouse, Nir Piterman, Lorenz Wernisch, Jasmin Fisher, and Berthold Göttgens. Btr: training asynchronous boolean models using single-cell expression data. BMC bioinformatics, 17(1):1–18, 2016.

[114] Aman Agrawal, Alec M Chiu, Minh Le, Eran Halperin, and Sriram Sankararaman. Scalable probabilistic pca for large-scale genetic variation data. PLoS Genetics, 16(5):e1008773, 2020.

[115] Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C Hicks. A systematic evaluation of single-cell rna-sequencing imputation methods. bioRxiv, 2020.

[116] Shawn C Baker, Steven R Bauer, Richard P Beyer, James D Brenton, Bud Bromley, John Burrill, Helen Causton, Michael P Conley, Rosalie Elespuru, Michael Fero, et al. The external rna controls consortium: a progress report. Nature methods, 2(10):731, 2005.

[117] FDA SEQC, Wangqiu Chen, Yongmei Zhao, Xin Chen, Xiaojiang Xu, Zhaowei Yang, Yingtao Bi, Vicky Chen, Jing Li, Hannah Choi, et al. A comprehensive multi-center cross-platform benchmarking study of single-cell rna sequencing using reference samples. bioRxiv, 2020.

[118] Tallulah S Andrews, Vladimir Yu Kiselev, Davis McCarthy, and Martin Hemberg. Tutorial: guidelines for the computational analysis of single-cell rna sequencing data. Nature Protocols, pages 1–9, 2020.

Supplementary

Design of five masking schemes

Given the three synthetic datasets without zero inflation, we use five masking schemes to introduce a varying number of non-biological zeros into each dataset. Since there is no consensus on the generation mechanism of non-biological zeros, we design the five masking schemes to reflect two fundamental questions: whether the occurrence of non-biological zeros (1) depends on the actual gene expression levels and/or (2) is gene-specific. As the five masking schemes cover the extreme answers to both questions (Fig. S1a), we expect that they together cover the unknown generation mechanism of non-biological zeros and would thus reveal the realistic effects of non-biological zeros on cell clustering and DE gene identification.

We provide a toy example to demonstrate the five masking schemes in Fig. S1b and summarize their technical details in Fig. S1c. In short, for a dataset with n non-zero counts, given a masking proportion p, all schemes would mask approximately np non-zero counts. However, the five schemes differ in masking which np non-zero counts, and they can be categorized in two ways corresponding to the two aforementioned questions.

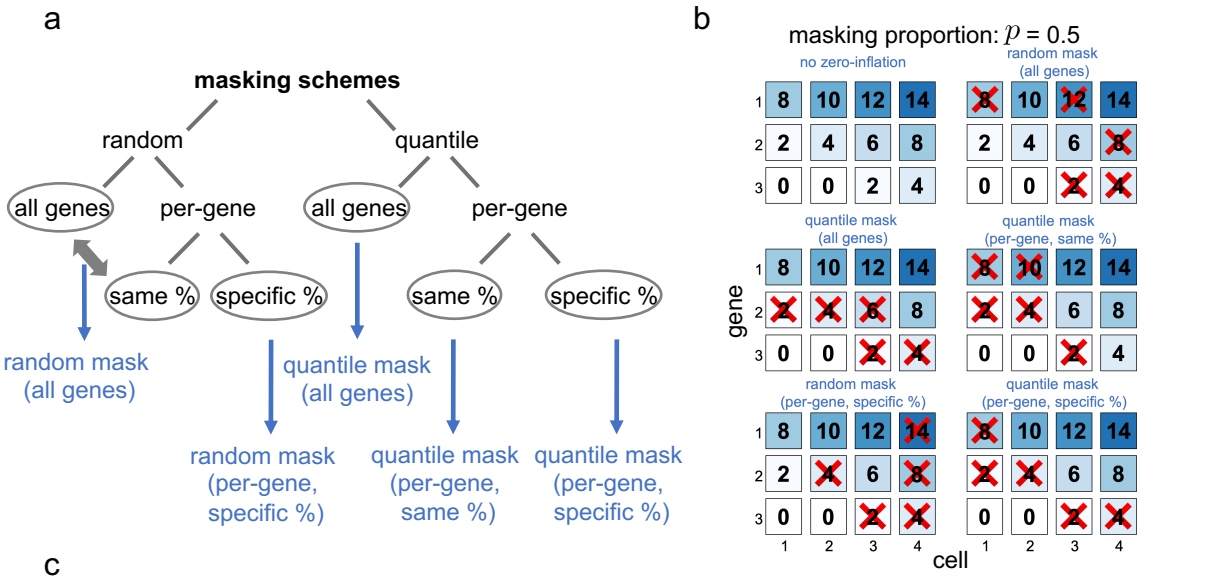
The first categorization is whether masking depends on the non-zero count values: random masking vs. quantile masking. While the two random masking schemes assume the independence between whether a non-zero count would be masked and the count value itself, the three quantile masking schemes assume a complete dependence by truncating non-zero values below a quantile (which corresponds to the masking proportion) to zero. Specifically, the two random masking schemes differ in the definition of independence: `random mask (all genes)` assumes the complete independence between masking and count values; `random mask (per-gene, specific %)` only assumes the conditional independence between masking and count values given each gene, and the masking proportion is gene-specific. Note that we define each gene's specific masking proportion as a function of the gene's non-zero counts based on an empirical formula in the literature [45, 46] (Fig. S1c); in short, the larger a gene's non-zero counts are, the smaller the gene's masking proportion is. Besides the two random masking schemes, the three quantile masking schemes differ in how they perform the truncation: `quantile mask (all genes)` truncates the lowest $100p\%$ non-zero counts of all genes; `quantile mask (per-gene, same %)` truncates the lowest $100p\%$ non-zero counts of each gene; `quantile mask (per-gene, specific %)` truncates the lowest non-zero counts of each gene based on the gene's specific masking proportion determined by the empirical formula.

The second categorization is regarding whether the masking proportion is gene-specific. Two schemes mask the same expected proportion $100p\%$ of non-zero counts for all genes: `random mask (all genes)` and `quantile mask (per-gene, same %)`. Three schemes use gene-specific masking proportions: `quantile mask (all genes)`, `random mask (per-gene, specific %)`, and `quantile mask (per-gene, specific %)`. Specifically, although `quantile mask (all genes)`

does not use the empirical formula to determine gene-specific masking proportions as in `random mask (per-gene, specific %)` and `quantile mask (per-gene, specific %)`, it still truncates different proportions of non-zero counts for different genes. The reason is that its truncation threshold is set to the p-th quantile of all genes' non-zero counts, and different genes have different numbers of non-zero counts below that threshold. It is also worth noting that we do not include `random mask (per-gene, same %)` because it is theoretically equivalent to `random mask (all genes)`—both schemes are expected to randomly mask $100p\%$ of every gene's non-zero counts (Fig. S1a).

Note that random masking aims to reflect the random nature of sampling zeros. In a sequencing experiment, allocation of reads to genes is essentially random sampling from a multinomial distribution, whose probabilities are the proportions of genes in terms of cDNA copy numbers in the sequencing library. Due to the randomness of sampling, for two genes with moderately different non-zero proportions, it is possible that, in one experiment, the gene with the larger proportion receives a zero read count, i.e., a sampling zero, while the gene with the smaller proportion receives a non-zero read count. The magnitude of the randomness depends on the sequencing depth. For every gene, the standard deviation of its count over its expected count is equal to a large constant depending on its proportion (i.e., $\sqrt{(1-q_i)/q_i}$, where q_i is the proportion of gene i) multiplied by the inverse of the square root of the sequencing depth (i.e., $1/\sqrt{N}$, where N is the sequencing depth). Hence, the smaller the sequencing depth, the larger the standard deviation of every gene's count in relation to its expected count, the more likely that genes receive sampling zeros irrespective of their proportions. Moreover, the expected number of sampling zeros (i.e., $\sum_{i=1}^{I}(1-q_i)^N$, where I is the number of genes) decreases as the sequencing depth increases. In contrast, quantile masking aims to reflect gene proportions in the sequencing library and technical zeros, i.e., zero counts due to zero proportions without randomness. Quantile masking also reflects the fact that, despite of randomness, a gene with a small proportion is more likely to receive a sampling zero than a gene with a much larger proportion does.

Hence, for Drop-seq and 10x Genomics, since they sequence many cells, per-cell sequencing depth is low and thus randomness is influential, random masking better represents the occurrence of non-biological zeros, sampling zeros in particular, than quantile masking does. For Smart-seq2, since per-cell sequencing depth is high and thus randomness is negligible, quantile masking better resembles the generation mechanism of non-biological zeros, technical zeros in particular, than random masking does.



c

masking scheme	parameter	method
random mask (all genes)	$p \in (0,1)$	randomly mask $N \sim$ Binomial(n,p) of the n non-zero counts of all genes
quantile mask (all genes)	$p \in (0,1)$	mask the smallest $\lfloor np \rfloor$ (floor) non-zero counts
qauntile mask (per-gene, same %)	$p \in (0,1)$	mask the smallest $\lfloor n_i p \rfloor$ non-zero counts of gene i, which has n_i non-zero counts
random mask (per-gene, specific %)	$\lambda > 0$, $p \in (0,1)$	randomly mask $N_i \sim$ Binomial(n_i, p_i) of the n_i non-zero counts of gene i, whose average (natrual) log-transformed non-zero count is μ_i, and $p_i = \exp(-\lambda\mu_i^2)$. λ is chosen to satisfy $\sum_{i=1}^{\#\text{ of genes}} n_i p_i = np$, where n is the total number of non-zero counts of all genes
quantile mask (per-gene, specific %)	$\lambda > 0$, $p \in (0,1)$	mask the smallest $\lfloor n_i p_i \rfloor$ non-zero counts of gene i. n_i and p_i are the same as in "random mask (per-gene, specific %)"

Figure S1: **Five masking schemes for introducing non-biological zeros. (a)** A tree diagram illustrating the design of the five masking schemes. From the top, the first division is about whether masking is independent of or completely dependent on count values, with the former as random masking and the latter as quantile masking. The second division is about whether masking is performed across all genes (with the same masking proportion) or within each gene (i.e., per-gene). If the latter, the third division is regarding whether the masking proportion is the same for all genes or specific to each gene depending on the gene's mean non-zero expression level. Note that random masking across all genes is equivalent to random masking per-gene with the same masking proportion (shown by the double arrow on the left). **(b)** A toy example illustration of the five masking schemes. The topleft plot shows the expression counts of three genes in four cells without zero-inflation; the other five plots show the expression counts after the five masking schemes are applied with the same masking proportion $p = 0.5$ (i.e., 50% of the non-zero gene expression counts are masked as zeros). **(c)** Technical explanation of each masking scheme. In the notations, p denotes the overall masking proportion across all genes, and p_i is the masking proportion of gene i.

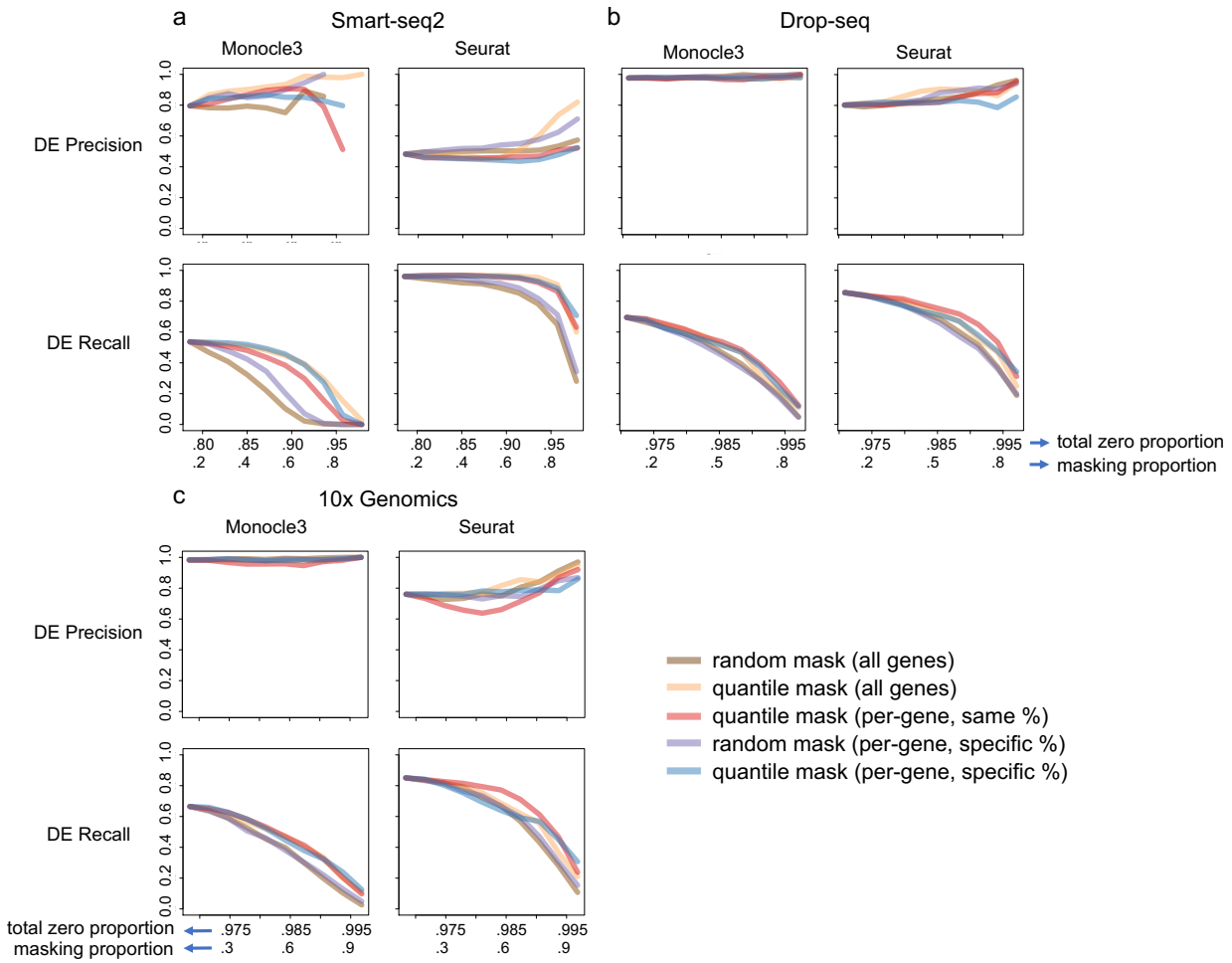


Figure S2: **Effects of non-biological zeros on DE gene identification in terms of precision and recall.** We introduce a varying number of non-biological zeros, which correspond to masking proportions 0.1–0.9, into the simulated **(a)** Smart-seq2, **(b)** Drop-seq, and **(c)** 10x Genomics datasets 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 schemes). After the introduction of non-biological zeros, we apply Monocle 3 and Seurat to each dataset to identify DE genes. We evaluate the accuracy using the precision and recall (given the false discovery rate 5%; defined in Fig. 6d), respectively.

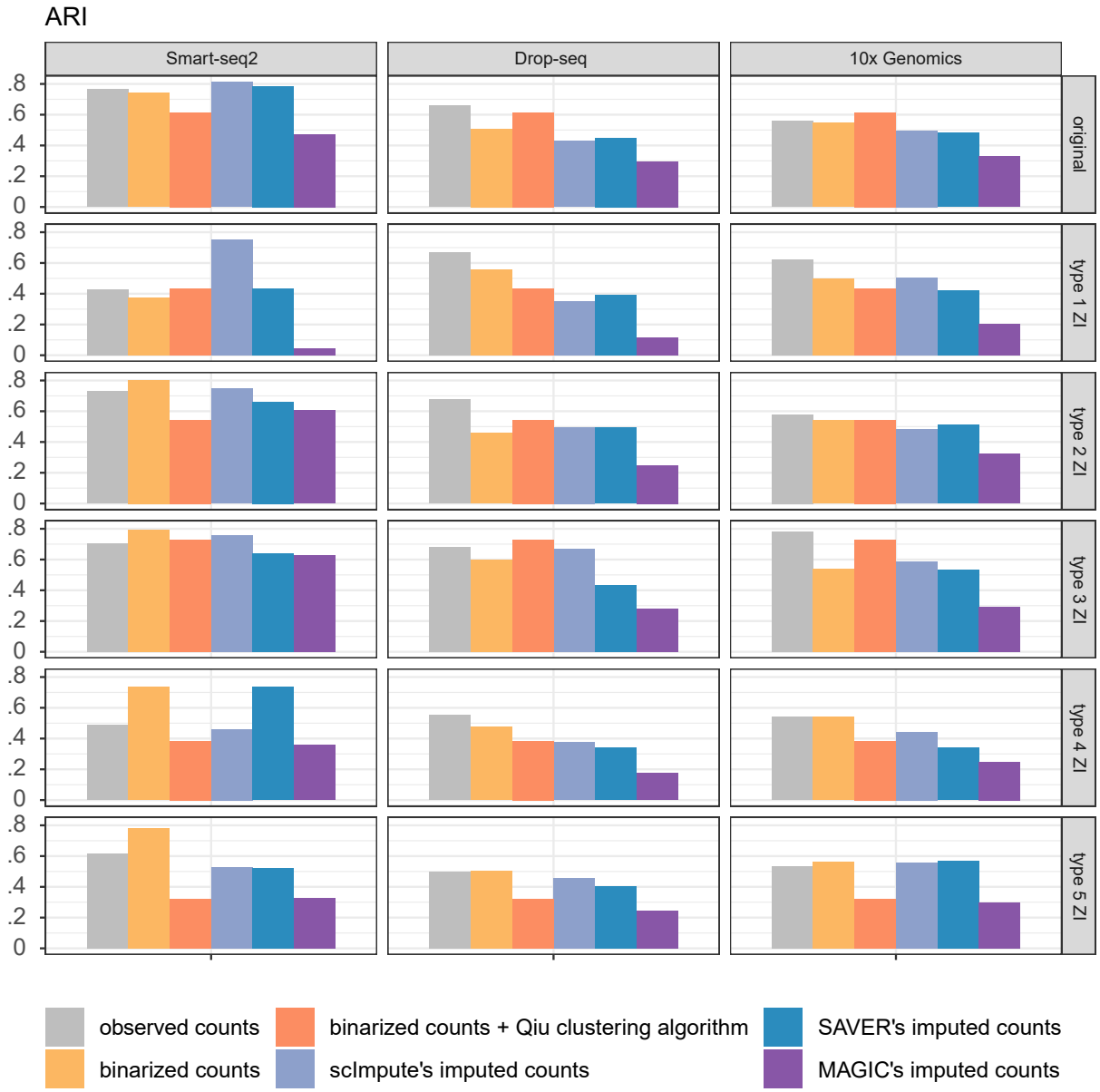


Figure S3: **Evaluation of clustering analysis on observed, binarized and imputed data.** We evaluate the clustering analysis on Smart-seq2, Drop-seq, and 10x Genomics data based on observed, binarized and imputed data. We perform this analysis before and after using the five masking schemes (type 1 ZI–type 5 ZI) to introduce non-biological zeros. Besides the bin-Qiu *et al.* which indicates the clustering algorithm developed specially for binarized data, we use Louvain clustering (in Seurat) on observed, binarized, and imputed data. We use ARI to evaluate the clustering results.

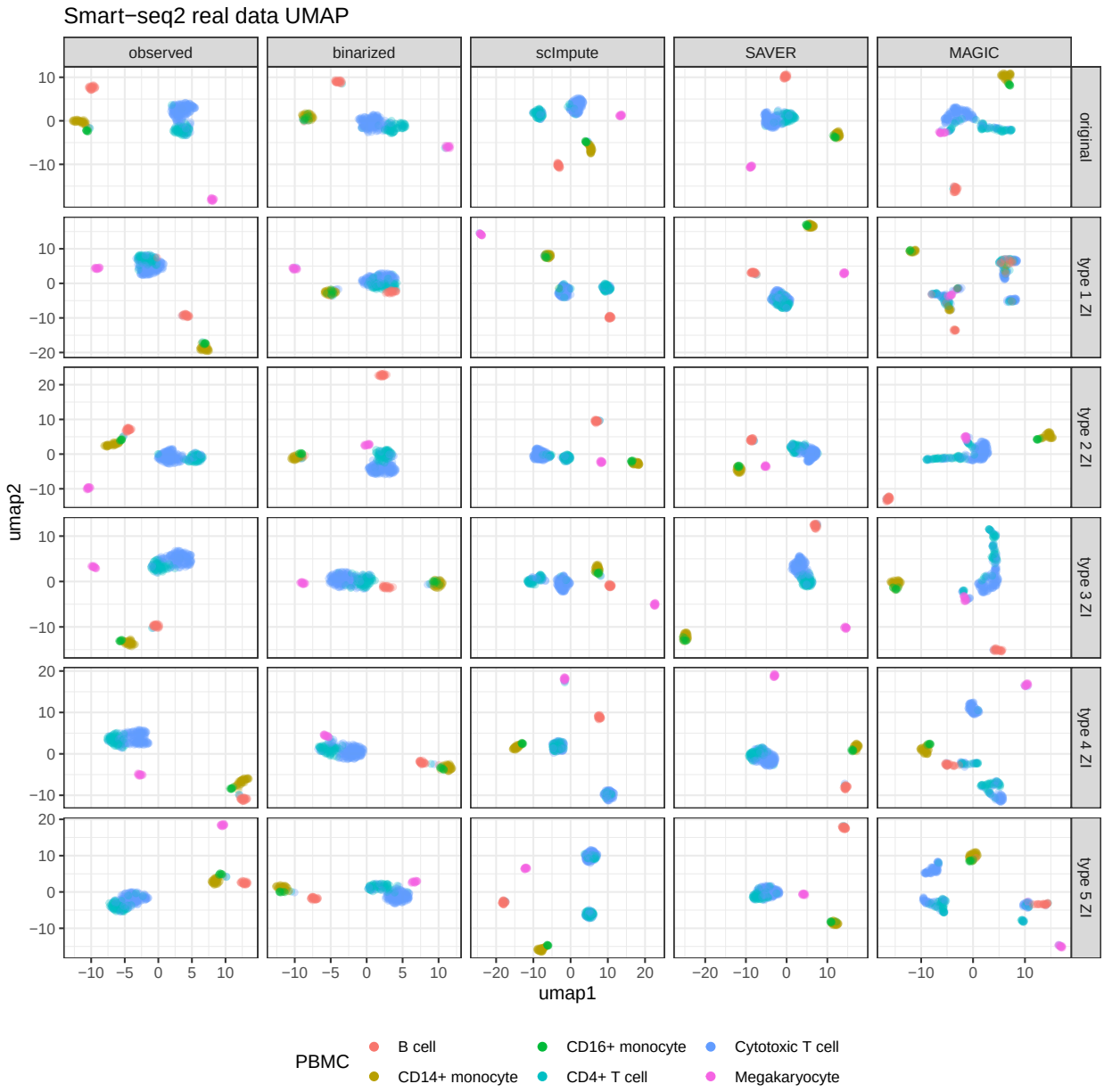


Figure S4: **UMAP dimesion reduction visualization on observed, binarized and imputed Smart-seq2 data.** We perform UMAP (in Seurat) on Smart-seq2's observed, binarized and imputed data. We perform this analysis before and after using the five masking schemes (type 1 ZI–type 5 ZI) to introduce non-biological zeros.

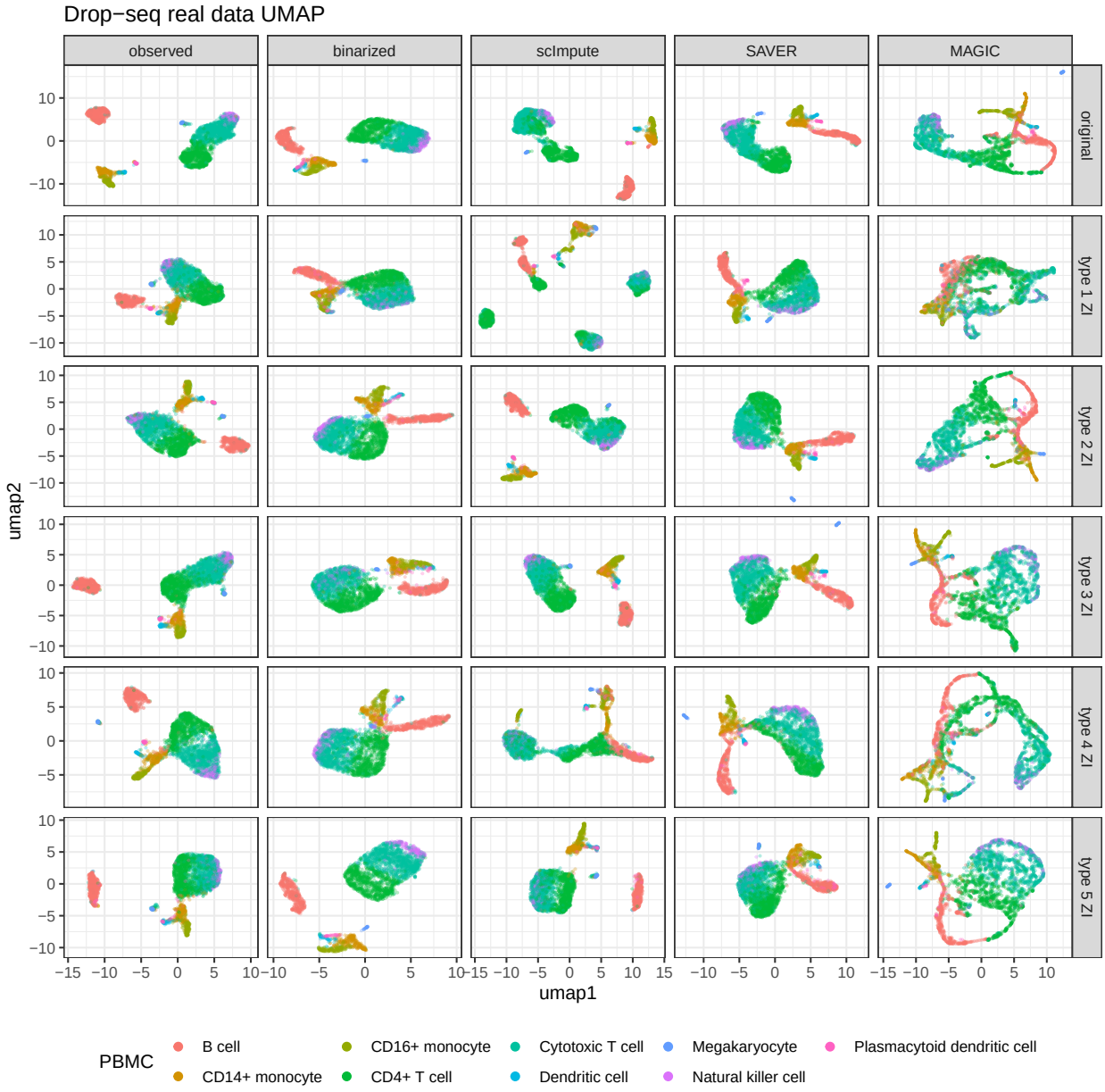


Figure S5: **UMAP dimesion reduction visualization on observed, binarized and imputed Drop-seq data.** We perform UMAP (in Seurat) on Drop-seq's observed, binarized and imputed data. We perform this analysis before and after using the five masking schemes (type 1 ZI–type 5 ZI) to introduce non-biological zeros.

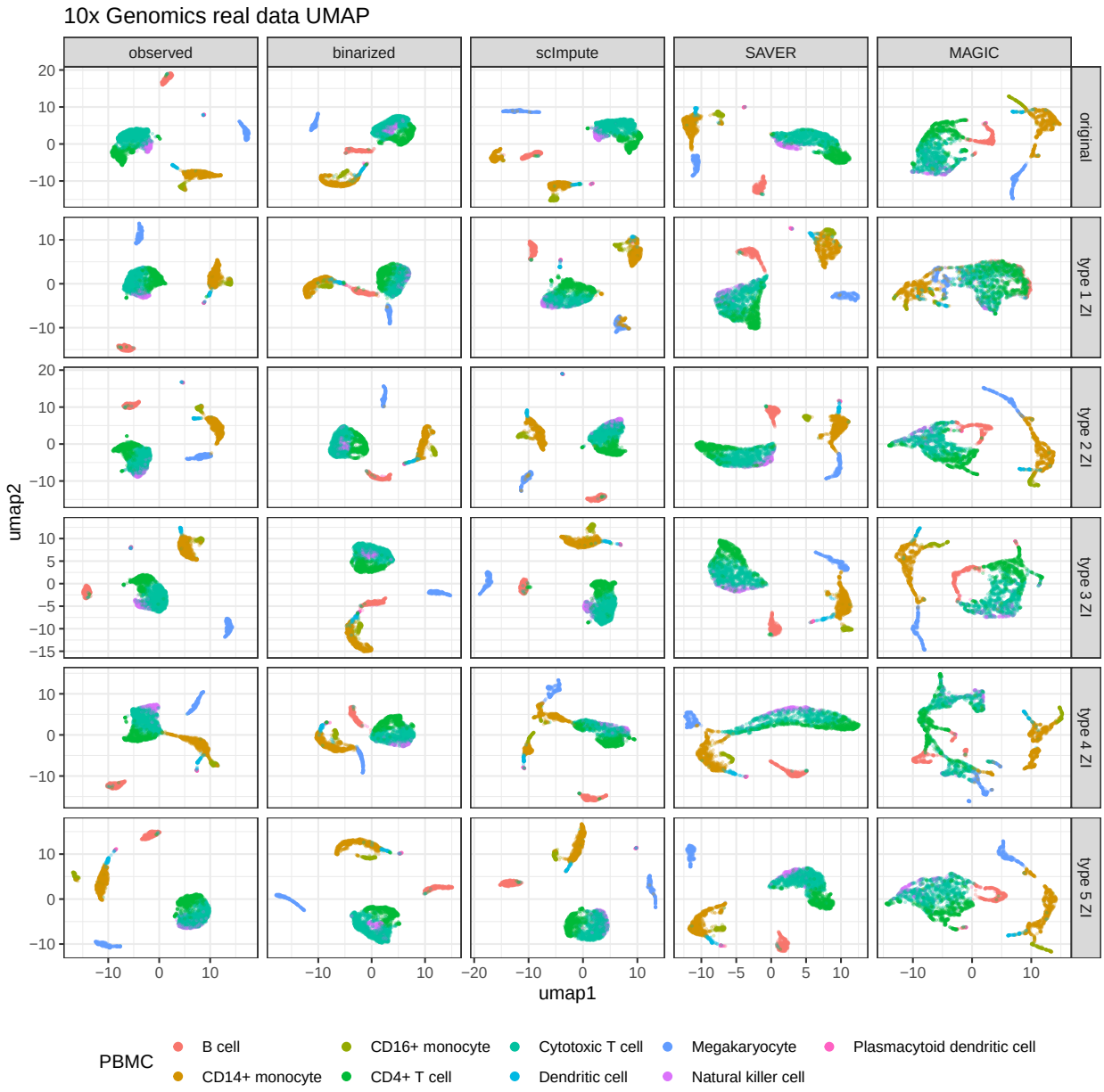


Figure S6: **UMAP dimesion reduction visualization on observed, binarized and imputed 10x Genomics data.** We perform UMAP (in Seurat) on 10x Genomics' observed, binarized and imputed data. We perform this analysis before and after using the five masking schemes (type 1 ZI–type 5 ZI) to introduce non-biological zeros.

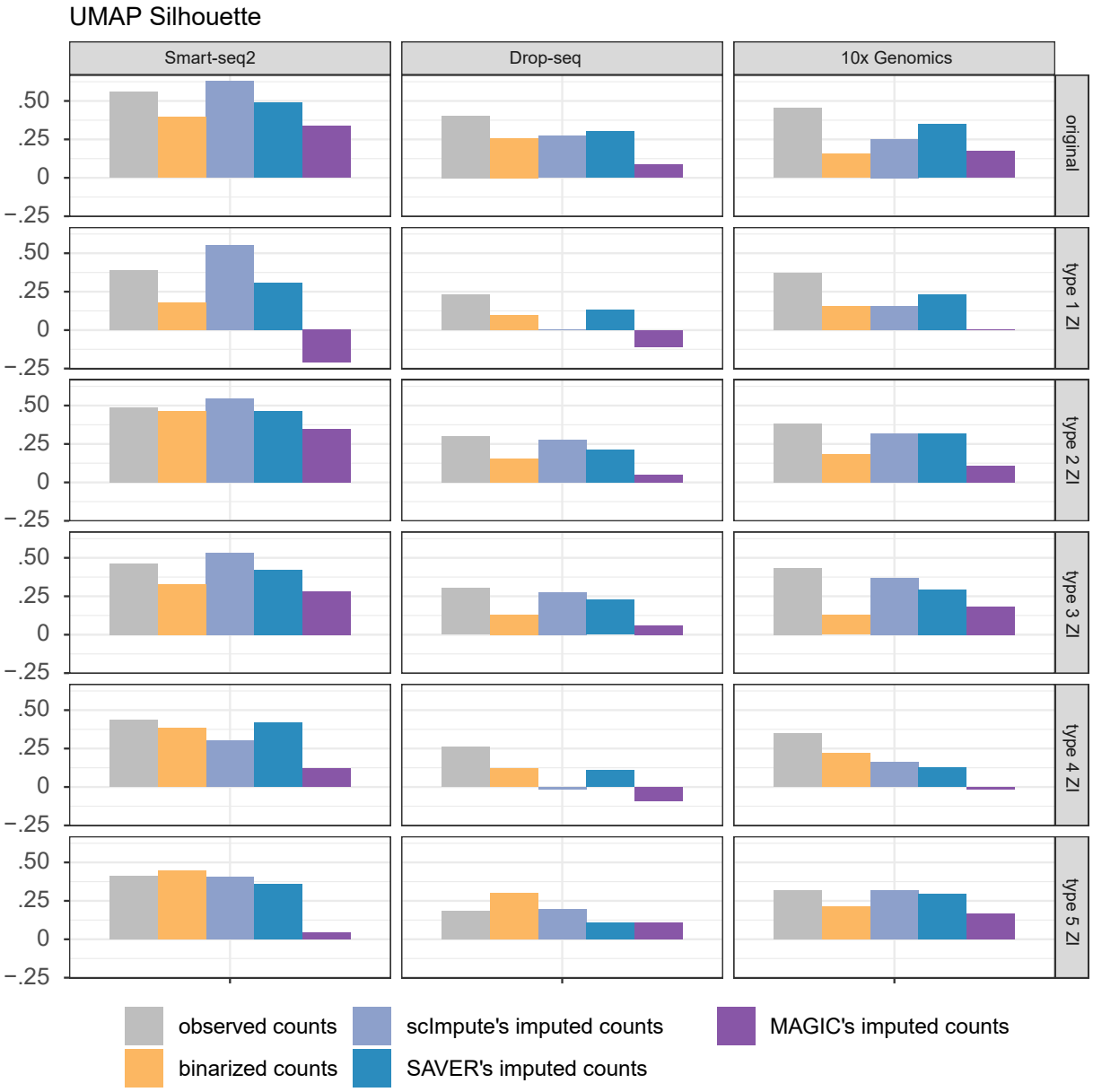


Figure S7: **Evaluation of dimension reduction analysis on observed, binarized and imputed data.** We evaluate the dimension reduction analysis on Smart-seq2, Drop-seq, and 10x Genomics data based on observed, binarized and imputed data. We perform this analysis before and after using the five masking schemes (type 1 ZI–type 5 ZI) to introduce non-biological zeros. We use UMAP (in Seurat) on observed, binarized, and imputed data to perform dimension reduction. We use Silhouette score to evaluate the dimension reduction results.

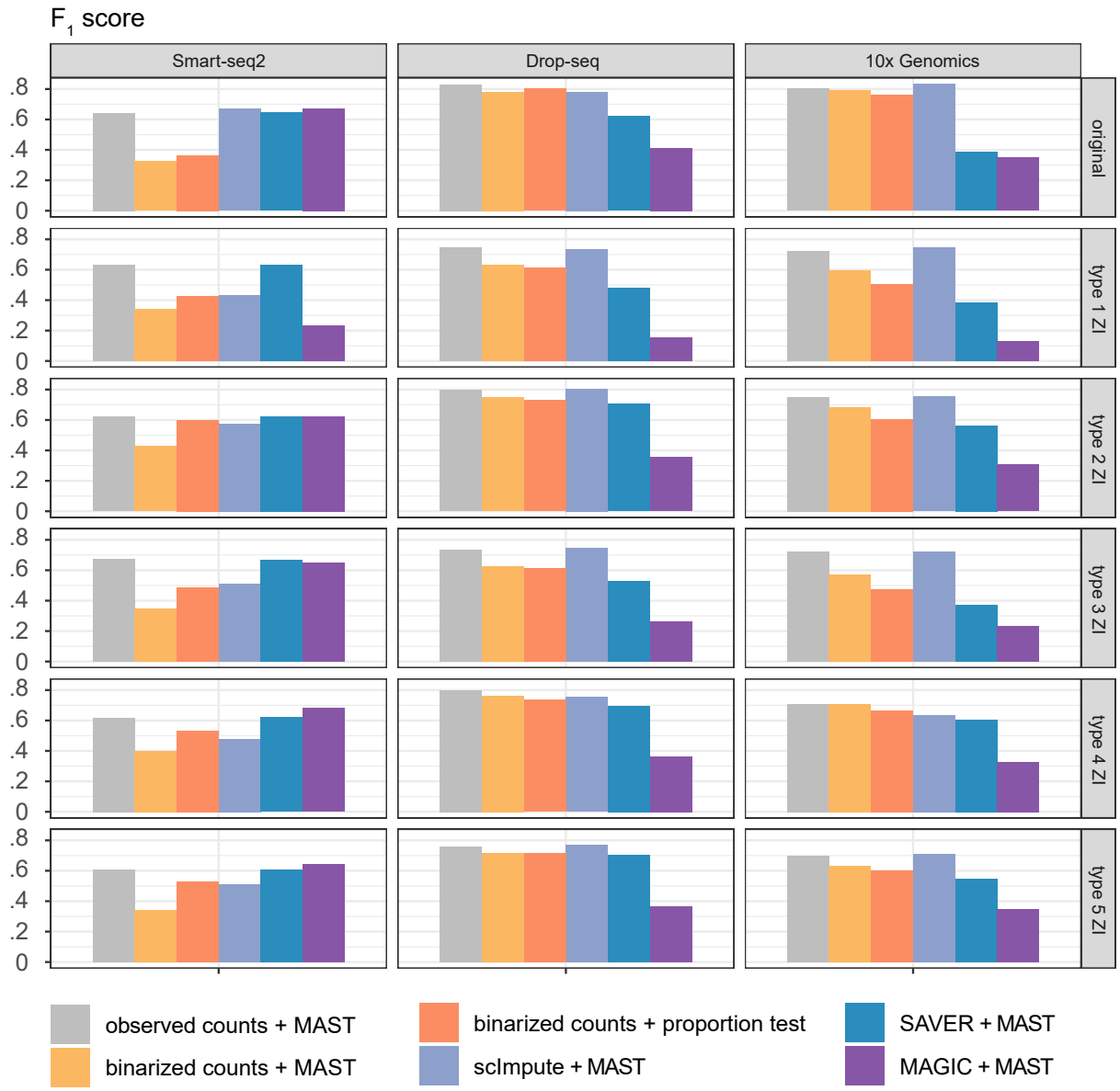


Figure S8: **Evaluation of DE analysis on observed, binarized and imputed data.** We evaluate the DE analysis on Smart-seq2, Drop-seq, and 10x Genomics data based on observed, binarized and imputed data. We perform this analysis before and after using the five masking schemes (type 1 ZI–type 5 ZI) to introduce non-biological zeros. We apply two-sample proportion test on binarzied data and MAST (in Seurat) on observed, binarized, and imputed data to perform DE analysis. We use F_1 score (given the false discovery rate 5%) to evaluate the DE results.

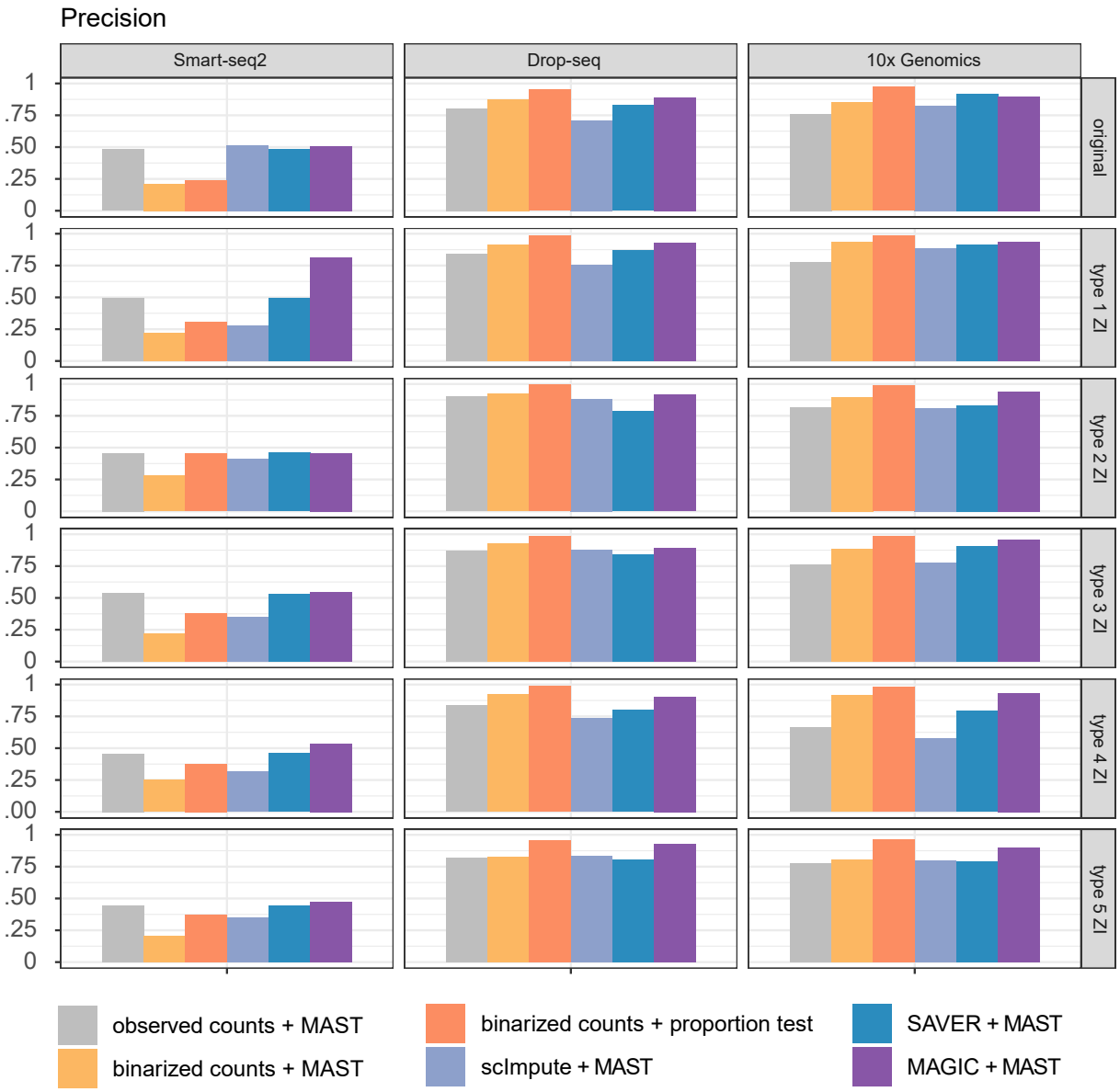


Figure S9: **Evaluation of DE analysis on observed, binarized and imputed data.** We evaluate the DE analysis on Smart-seq2, Drop-seq, and 10x Genomics data based on observed, binarized and imputed data. We perform this analysis before and after using the five masking schemes (type 1 ZI–type 5 ZI) to introduce non-biological zeros. We apply two-sample proportion test on binarzied data and MAST (in Seurat) on observed, binarized, and imputed data to perform DE analysis. We use precision (given the false discovery rate 5%) to evaluate the DE results.

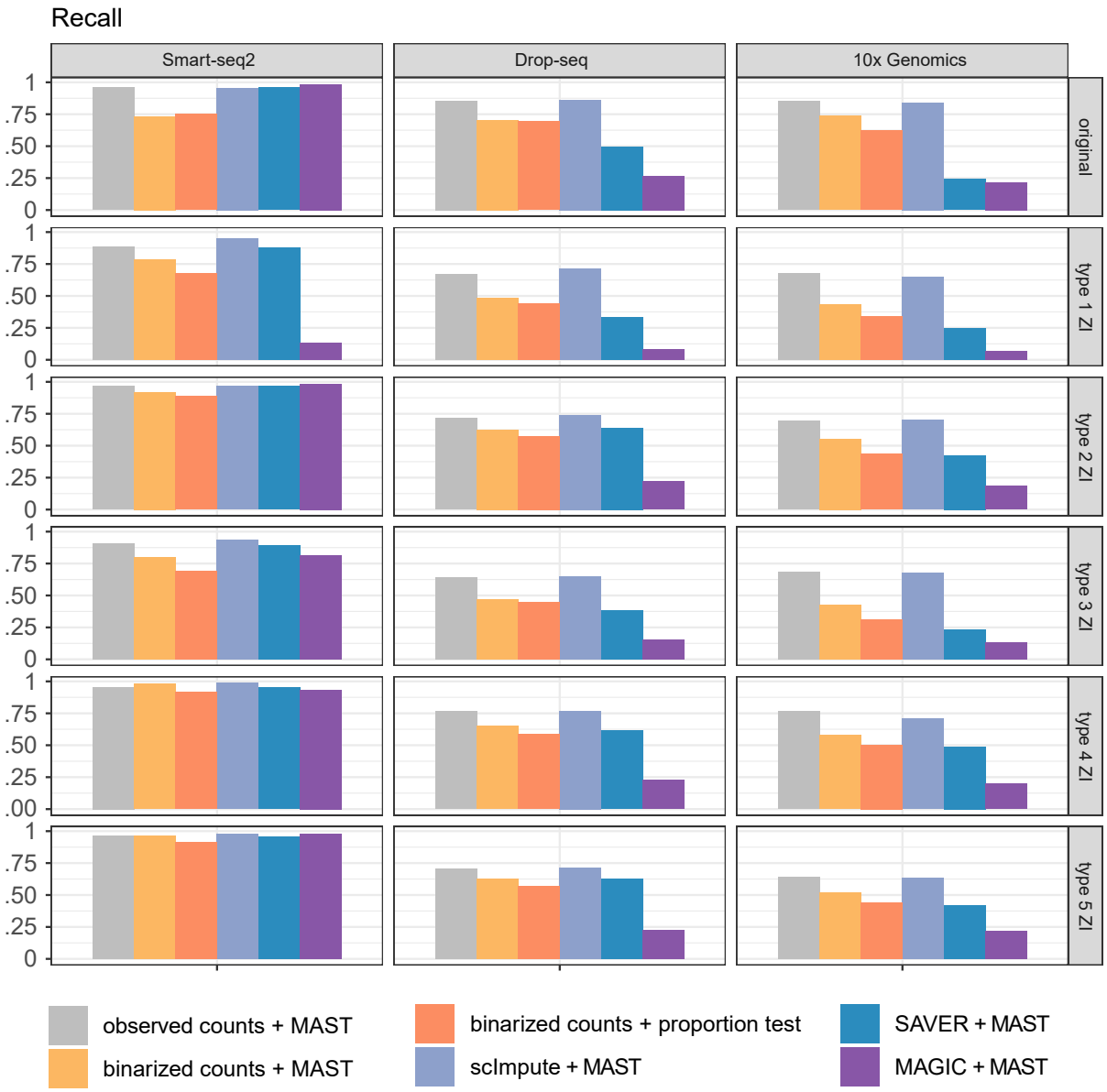


Figure S10: **Evaluation of DE analysis on observed, binarized and imputed data.** We evaluate the DE analysis on Smart-seq2, Drop-seq, and 10x Genomics data based on observed, binarized and imputed data. We perform this analysis before and after using the five masking schemes (type 1 ZI–type 5 ZI) to introduce non-biological zeros. We apply two-sample proportion test on binarzied data and MAST (in Seurat) on observed, binarized, and imputed data to perform DE analysis. We use recall (given the false discovery rate 5%) to evaluate the DE results.