Computational and Structural Biotechnology Journal 18 (2020) 2789–2798

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Naught all zeros in sequence count data are the same

Justin D. Silverman [a,b,c], Kimberly Roche [d], Sayan Mukherjee [d,e,f,*], Lawrence A. David [d,f,g,*]

[a] College of Information Science and Technology, Pennsylvania State University, State College, PA 16802, United States
[b] Institute for Computational and Data Science, Pennsylvania State University, State College, PA 16802, United States
[c] Department of Medicine, Pennsylvania State University, Hershey, PA 17033, United States
[d] Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27708, United States
[e] Departments of Statistical Science, Mathematics, Computer Science, Biostatistics & Bioinformatics, Duke University, Durham, NC 27708, United States
[f] Center for Genomic and Computational Biology, Duke University, Durham, NC 27708, United States
[g] Department of Molecular Genetics and Microbiology, Duke University, Durham, NC 27708, United States

ARTICLE INFO

Article history:
Received 26 June 2020
Received in revised form 9 September 2020
Accepted 10 September 2020
Available online 28 September 2020

Keywords:
Sequence count data
Microbiome
Statistics
Gene expression
Zero counts

ABSTRACT

Genomic studies feature multivariate count data from high-throughput DNA sequencing experiments, which often contain many zero values. These zeros can cause artifacts for statistical analyses and multiple modeling approaches have been developed in response. Here, we apply different zero-handling models to gene-expression and microbiome datasets and show models can disagree substantially in terms of identifying the most differentially expressed sequences. Next, to rationally examine how different zero handling models behave, we developed a conceptual framework outlining four types of processes that may give rise to zero values in sequence count data. Last, we performed simulations to test how zero handling models behave in the presence of these different zero generating processes. Our simulations showed that simple count models are sufficient across multiple processes, even when the true underlying process is unknown. On the other hand, a common zero handling technique known as "zero-inflation" was only suitable under a zero generating process associated with an unlikely set of biological and experimental conditions. In concert, our work here suggests several specific guidelines for developing and choosing state-of-the-art models for analyzing sparse sequence count data.

Contents

* Co-Corresponding author.

https://doi.org/10.1016/j.csbj.2020.09.014

1. Introduction

Many high-throughput DNA sequencing assays exhibit high sparsity, which often exceed 70% in microbiome, bulk-, and single-cell RNA-seq experiments [1–4]. Such sparsity can be problematic for modeling [5,3,6,7], as common numerical operations like logarithms or division are undefined when applied to zero. Empirical benchmarks also suggest that the frequency of zeroes in datasets can affect false discovery rates in analyses like differential gene expression [8,9].

Multiple approaches have been proposed for tackling the modeling problems posed by zero values in sequence count data. A common approach for addressing numerical challenges associated with taking the logarithm or dividing by zero is to add a small positive value, or pseudo-count, to the entire dataset prior to analysis [10,5]. A more sophisticated approach is to model all counts (including zero values) as arising due to random counting involving the Poisson, negative binomial, or multinomial distributions [11–15]. Often these methods perform inference on the statistical properties of the entire datasets rather than a single observed zero count. Still more complicated models permit greater flexibility in the modeling of zero values by layering secondary random processes on top of random count processes. Examples of such models include zero-inflated negative binomial models [16] and Poisson zero-inflated log-normal models [17].

Although an abundance of methods have been proposed for handling zeros, it remains unclear when certain approaches are to be preferred over others. Empirical benchmarks comparing sequence analysis software packages [8,9] do not isolate the effects of zero handling relative to other modeling decisions such as how to filter samples or normalize read depths, which in turn precludes offering specific guidance as to which approaches to modeling zero values are more or less appropriate. A conceptual debate has also emerged around the appropriateness of zero-inflation, with some arguing it is unnecessary [18,19,17], while others have suggested it can account for the large number of zero values in single-cell RNA-seq data [20–25,16,26–28,4], bulk RNA-seq data [29–32], and microbiome sequencing data [33–42]. This debate reflects controversy as to what kinds of processes give rise to zero values in sequence count data [18,19,17,20].

In this Perspective, we re-analyze published datasets to show that alternative methods of zero–handling can lead to different inference outcomes. To understand the origins of these differences, we introduce a categorization scheme for zero generating processes (ZGPs) in sequence count data. While the precise ZGPs that contributed to a given dataset are typically unknown, we can use simulation to examine if there exist zero–handling models that perform well across a range of different ZGPs. Overall, our analyses reveal minimal conceptual and analytical support for the use of zero-inflated models to handle zeros in sequencing datasets. Our results suggest that simpler models avoiding zero-inflation are preferable for most tasks.

2. Real data examples

To investigate whether different methods of modeling zero values can affect the outcomes of real-world analyses, we reanalyzed six previously published datasets using models that differed only in their handling of zero values; one model used zero-inflation while the other did not. We chose datasets that spanned a range of sequencing tasks: single cell RNA-seq [43,44], bulk RNA-seq [45,46], and 16S rRNA microbiota surveys [47,48]. We then chose two different statistical models that differed only in their modeling of zero values. One model was based on a negative binomial distribution. Letting y_{ij} represent the observed counts for sequence i in sample j, a negative binomial model assumes that y_{ij} reflects the abundance of sequence i in sample j with added sampling noise described by a negative binomial distribution. Such negative binomial models are used in many popular software tools such as edgeR [49], DESeq2 [11]. The second model was similar to the first, but also assumed a process known as zero-inflation was taking place. In contrast to the first model, a *zero-inflated* negative binomial model additionally assumes that there exists a probability π_j that $y_{ij} = 0$ regardless of the abundance of sequence i in sample j. Zero inflation has become a popular method of augmenting the negative binomial to model higher levels of zero values in sequence count data [29,20,38,16,28]. To implement our two models, we used the ZINB-WaVE modeling framework [16], which allowed us to create identical negative binomial models that varied only according to the presence of zero inflation. Notably, our implementation relied on the default settings of ZINB-WaVE, which further assumes that the probability π_j can vary depending on the condition a sample belongs to (e.g. treatment or control). Such condition-specific zero-inflation is commonly used in a number of popular software packages [33,20,25,16,27]. Moreover, the zero-inflated model we use here has been used in multiple studies to conduct differential expression analysis [50–53]. We refer to these two models respectively as the Zero-Inflated Negative Binomial (ZINB) and Negative Binomial (NB) models (see Section 6 for more details).

To interpret the results of these two models we quantified the discrepancy between the top-K most differentially expressed sequences according to each model (Figure S1). Discrepancy was calculated as $(K - m)/K$ where m is the number of the top-K sequences in common between the two models. We found that the ZINB and NB models disagreed on average by 44% (range: 14%-100%) among the top-50 most differentially expressed sequences. Even among the top-5 most differentially expressed sequences (a subset whose size and priority makes them likely for potentially costly experimental follow-up), disagreement averaged 53% and reached 100% for one dataset (Fig. S1).

We found that the largest discrepancies between the ZINB and NB models occurred on sequences that were observed with a high number of counts in one condition, while also being observed with low or zero counts in the other condition (Fig. 1, S2, and S3). These presence-absence-like cases would seem like examples of where sequence abundance varies according to condition, and indeed, the NB model infers these sequences are differentially expressed (Fig. 1, S2, and S3). By contrast, we observed that the ZINB model does not always infer that sequences exhibiting presence-absence-like patterns in one condition were differentially expressed. The ZINB model instead inferred that these sequences were actually expressed at equal abundance; but, one condition exhibited higher rates of zero-inflation than the other condition (a phenomenon we term differential zero-inflation; Fig. 1, S2, and S3). Indeed, we found that there was a correlation between the difference in inferred differential expression according to the ZINB and NB models and the degree to which the ZINB model inferred that a sequence is differentially zero-inflated (Spearman $\rho > 0.35$ and p-value ≈ 0 for all 6 datasets; Fig. 1, S2, and S3). Still, discrepancies between how the NB and ZINB models handled sequences with presence-absence like abundance patterns were not solely due to condition-specific zero-inflation, as we could recapitulate similar levels of discrepancy (average of 37%; range: 12%–86%; Figs. S4-S7) even when using non-condition-specific zero-inflation (see Section 6 and Supplementary File 1 for complete discussion and methods). Ultimately, while the true processes underlying sparsity in a genomic dataset are often unknown, we found it striking that sequences with high counts in one condition, while also being observed with low or zero counts in the other condition, were often inferred by the ZINB model to not be differentially expressed. This

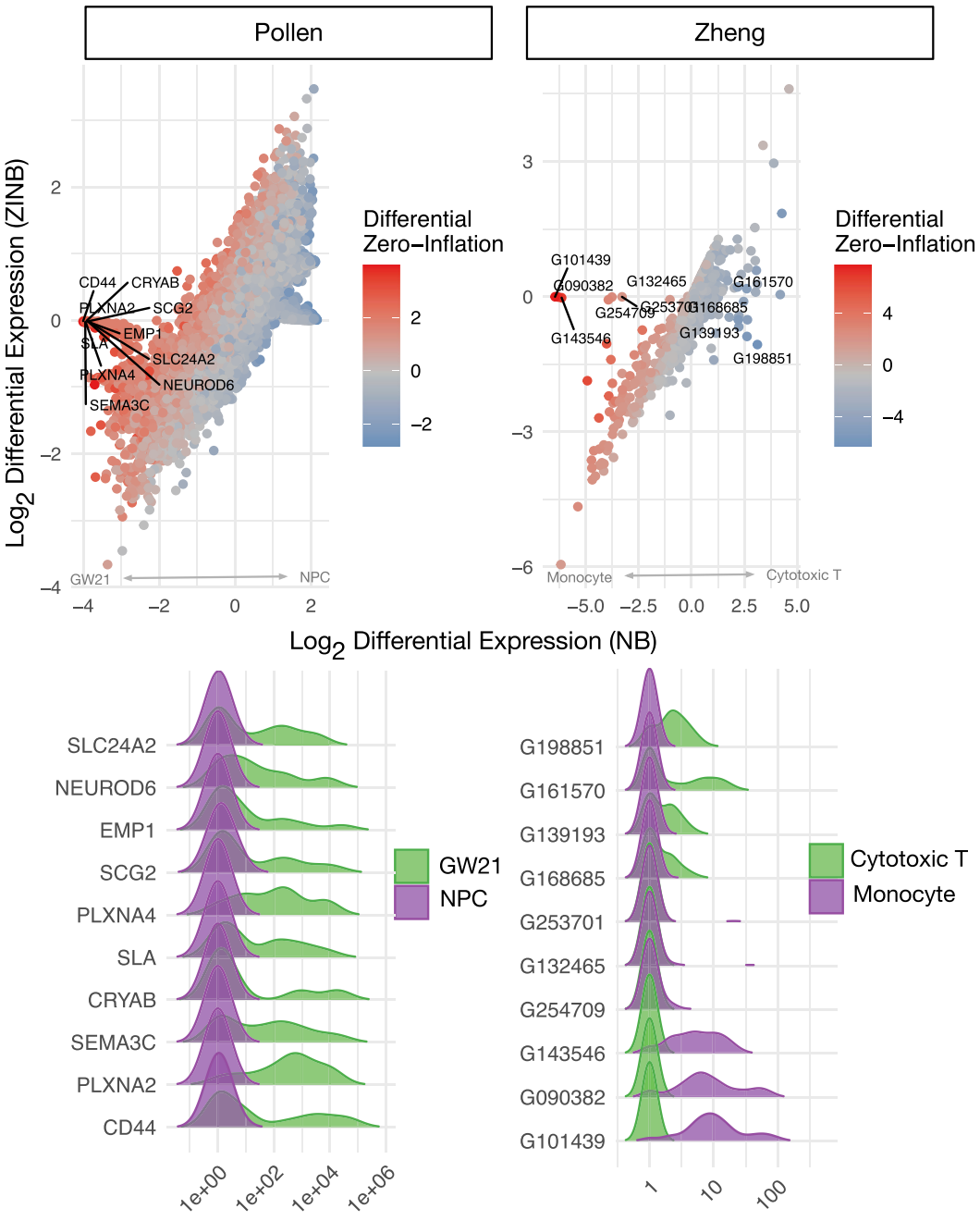

Fig. 1. Differential expression (DE) estimates from a negative binomial (NB) and zero-inflated negative binomial (ZINB) model can differ substantially. Log base 2 differential expression for the ZINB and NB models are shown after each was applied to two different single cell RNA-seq datasets. The orientation of differential expression is denoted by an arrow above the X axis; for example, in the Pollen dataset, genes higher in the NPC condition correspond to larger values of differential expression. Dots represent different genes, and each is colored according to the degree of differential zero-inflation as estimated by the ZINB model. For each dataset, the 10 genes that have the largest discrepancy between inferred DE are labeled and their distribution is in each condition is plotted in the bottom panel. Similar figures for two bulk RNA-seq datasets and two 16S rRNA surveys are shown in Figures S2 and S3 respectively.

suggest that zero-inflated models lead to higher false-negative rates than identical non-zero-inflated models.

3. Zero Generating Processes (ZGPs)

To provide a conceptual framework for analyzing how different zero handling models behave, we developed a scheme for categorizing different zero generating processes (ZGPs). Our scheme partitions ZGPs into three major classes, one of which we further subdivide into two sub-classes (Fig. 2):

Sampling Zeros: Zeros may also arise due to limits in the total number of sequencing reads counted in a given sample [3]. Certain sequences, particularly ones at low abundance, may be present but not counted. In the limit where no reads were collected for a given sample, all zeros would be due to sampling effects; by contrast, if every read in a sample could be collected for a sample, sampling zeros would not be present.

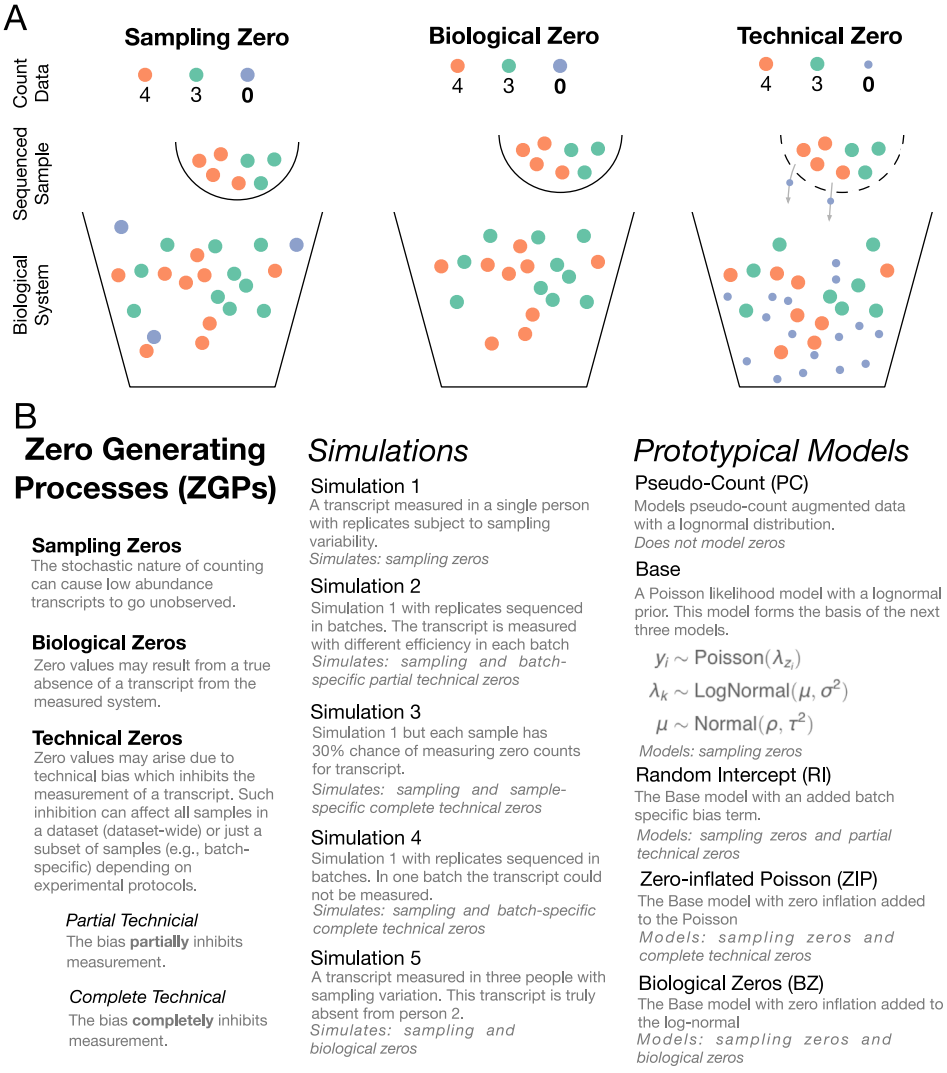

Fig. 2. An overview of the zero generating processes (ZGPs), simulations, and models presented in this work. (A) A graphical representation of three major ZGPs. Orange, green and blue represent distinct DNA sequences. (B) Model notation is as follows: y_i represents the number of counts observed for a given sequence in sample i, z_i represents the person from which sample i originates, x_i represents the batch number of sample i, λ_{z_i} represents the abundance of the sequence in person z_i. σ^2, ρ, and τ^2 are fixed hyper-parameters of the model.

Biological Zeros: Perhaps the most intuitive reason for zeros in a dataset, biological zeros arise when a sequence is truly absent from a biological system.

Technical Zeros: Preparing a sample for sequencing can introduce technical zeros into the data by **partially** or **completely** reducing the amount of countable sequences. These processes can lead to reductions in sequence abundance across all samples in a study or can act on a subset samples. For example, some genes are under-represented (partially reduced) in sequencing libraries due to the relative difficulty of amplifying GC-rich sequences [54,55]. This bias could occur across all samples in a study; or, in a heterogeneous manner if different batches of samples were amplified using different primers or cycle number. Furthermore, if instead of a relative inability to amplify a sequence there was a *complete* inability, we would have a complete technical process. Zero-inflated models consider a specific case of complete technical zeros where this inability to measure a given sequence occurs randomly from sample-to-sample. Importantly, in a complete technical process, even abundant sequences may go unobserved.

There is ample experimental evidence that biological, sampling, and partial technical zeros occur in real data. Biological zeros are known to occur when studying gut microbiota across people [56] since unrelated individuals will harbor unique bacterial strains. Another example of biological zeros can be found in RNA expression analyses of gene knockout experiments, where gene deletion will eliminate certain transcripts from the expressed pool of genes prior to DNA sequencing [57]. Sampling zeros are known to occur when sequencing depth is limited and sequence diversity is high [2,18]. For example, *in silico* studies have shown that decreasing the depth of sequencing studies can increase the numbers of observed zeros [36]. There also exist well-known examples of partial technical processes such as DNA extraction bias [58], batch effects [59,60] or PCR bias [55,61–63].

In contrast with the other ZGPs, the rationale for modeling complete technical zeroes is more nuanced. Dataset-wide complete technical zeros certainly exist: certain prokaryotic taxa are known to not be detectable by common 16S rRNA primers, for example, and will therefore not appear in microbiota surveys [64,65]. Of course, modeling the expression of dataset-wide technical zeros is typically neither considered a useful nor practical endeavor. Much more interest and effort though has been invested in considering cases of sample-specific complete technical zeros; that is,

when with some probability π_j, a sequence j may go completely unobserved in a given sample, regardless of its true abundance. Such sample-specific complete technical zeros have been considered in the analysis of gene-expression data, where zero-inflation has previously be used to model a variety of processes often termed "dropout" [20,22,23,25,27]. Dropout has been explained as stochastic forces involved with sampling of low-abundance sequences or due to the stochastic nature of gene expression at the single-cell level [20,22]. Dropout has also been described as the result of failures in amplification during the reverse-transcription step in RNA-seq experiments [20,22]. In the analysis of microbiome data, zero-inflation has been used to model differing presence/absence patterns between individuals [20,22]. Yet, each of the aforementioned phenomena could be argued as arising from either sampling, biological, or partial technical ZGPs. The inability to detect low abundance RNA sequences could be argued as arising from a sampling process rather than a complete technical process. Stochastic gene expression at the single-cell level and differing presence/absence patterns between individuals fit the definition of biological zeros. Last, zeros associated with gene amplification inefficiency could either reflect a sampling process or a partial technical process related to competitive inhibition between sequences in the amplification reaction. Thus, models considering sample-specific complete technical processes may actually be attempting to capture other ZGPs.

4. Simulation studies

As the ZGPs present in a given dataset are typically unknown, we sought to understand how different methods of handling zeros performed under model mis-specification. Exploring model behavior using empirical benchmarks that compare common software platforms (e.g. DESeq2, EdgeR, or ALDEx2) is not optimal because multiple modeling decisions independent of zero–handling like data filtering, inference method, and data normalization are incorporated into commonly used sequence analysis tools. To isolate the effects of different zero–handling methods on sequence analysis, we created a 5 by 5 grid in which different models were tested against different combinations of ZGPs. We started with a simple inference model based on the Poisson distribution (Base model, Fig. 2). We chose this distribution in place of the negative binomial (also called the Poisson-Gamma) distribution (which was used in our analysis of real data) because the Poisson affords a simpler to understand, single-parameter model while still modeling random sampling. To understand how the Poisson model interpreted the data we used tools from Bayesian statistics which allowed us to directly calculate a probability distribution representing the models belief in a sequence's abundance having observed sequence count data (the posterior distribution) [66]. To accomplish this, we also needed to specify a distribution representing the model's belief in the abundance of the sequence prior to seeing the data (prior distribution) [66]. We chose the log-normal prior in place of the more common gamma distribution because the mean and variance parameters of the log-normal distribution provide a more interpretable description of sequence abundances than the shape and rate parameters of the gamma distribution.

We then layered four common zero-modeling approaches onto our Base model (Fig. 2; full descriptions of each model are presented in *Methods*). We created a Zero-Inflated Poisson (ZIP) model by adding a zero-inflation component to the Poisson distribution in the Base model. To model zeros arising due to true biological absence, we created the Biological Zero model (BZ). This model includes a zero inflated component on the Log-Normal portion of the Base model and is similar to the DESCEND model of Wang et al. [17]. To model sampling and partial technical zeros, we created the Random Intercept (RI) model. The RI model differs from the Base model only by allowing external covariate information, *e.g.*, batch numbers, to decrease the amount of available sequences. Last, we designed a Pseudo-Count (PC) model that does not model any ZGPs, but rather avoids numerical issues with zeros by adding a fixed positive pseudo-count κ to each observed count value (i.e., each cell) in the data before analysis. The PC model removes the Poisson component from the Base model and instead uses $\hat{\lambda}_i = y_i + \kappa$ as the observation.

We next designed five different simulation experiments to test our hypothesis that some zero–handling methods would be more tolerant to model mis-specification than others. As sampling lies at the heart of sequence count data [67,18], we included zeros from Poisson sampling in each simulation. In addition, to investigate the other ZGPs, the second through fifth simulations included biological, complete-, and partial-technical processes as well. Each model was applied when simulations could distinguish models. For example, in the absence of covariate information the RI and Base models are identical; therefore, the RI model was only applied to simulations where covariate information was available and could distinguish the RI and Base models (*See Supplementary File 1 for a full discussion of this topic*).

We judged model performance on a given simulation based on the inferred probability that the sequence's abundance was less than or equal to the true simulated value, *i.e.*, the cumulative distribution function of the posterior density evaluated at the true value. This statistic captures both the error in the model's best guess (the mean) as well as its certainty about that guess (the spread of the posterior about the mean; Figure S8 provides a visual explanation of this metric). An optimal model would have a value of 0.5, whereas models that performed poorly would have values near 0 or 1 if they under- or over-estimated the true value with undue certainty.

The results of our simulation studies are shown in Figure 3 and S9. (A detailed description and explanation of the results is presented in *Supplementary File 1*.) Overall, we found the best-performing model to be the RI model. In particular, this model displayed three beneficial features. First, like the Base model, the RI model appropriately modeled sampling zeros without difficulty. Second, even though it did not model biological zeroes directly, the RI model approximated biological zeros as very low abundance; conveying the key information, that the absent sequence was not common. Third, by allowing covariate adjustment, the RI model effectively estimated sequence abundance even in the presence of partial technical zeros due to batch effects.

The next best performing models were the Base and BZ models. The Base model exhibited similar simulation results as the RI model, with the exception of its performance on technical zero simulations due to its inability to incorporate additional covariates like batch. The BZ exhibited a similar limitation on modeling external covariates and hence, batch effects. Still, the BZ model performed well on modeling of biological zeroes, which was expected because it is designed specifically for such a phenomenon.

More poorly performing models in our simulations were the PC and ZIP models. In addition to its observed lower accuracy, the PC model was also sensitive to chosen pseudo-count values (Fig. 3A and E). The ZIP model overestimated sequence abundances (λ) in every simulations with the exception of the one explicitly simulating sample-specific complete technical zeroes (Simulation 3). Additionally, our simulations showed that the ZIP model had high posterior uncertainty: The ZIP model could not tell if zeros were due to sampling of a low abundance sequence, technical censoring of a low abundance sequence, or technical censoring of a high

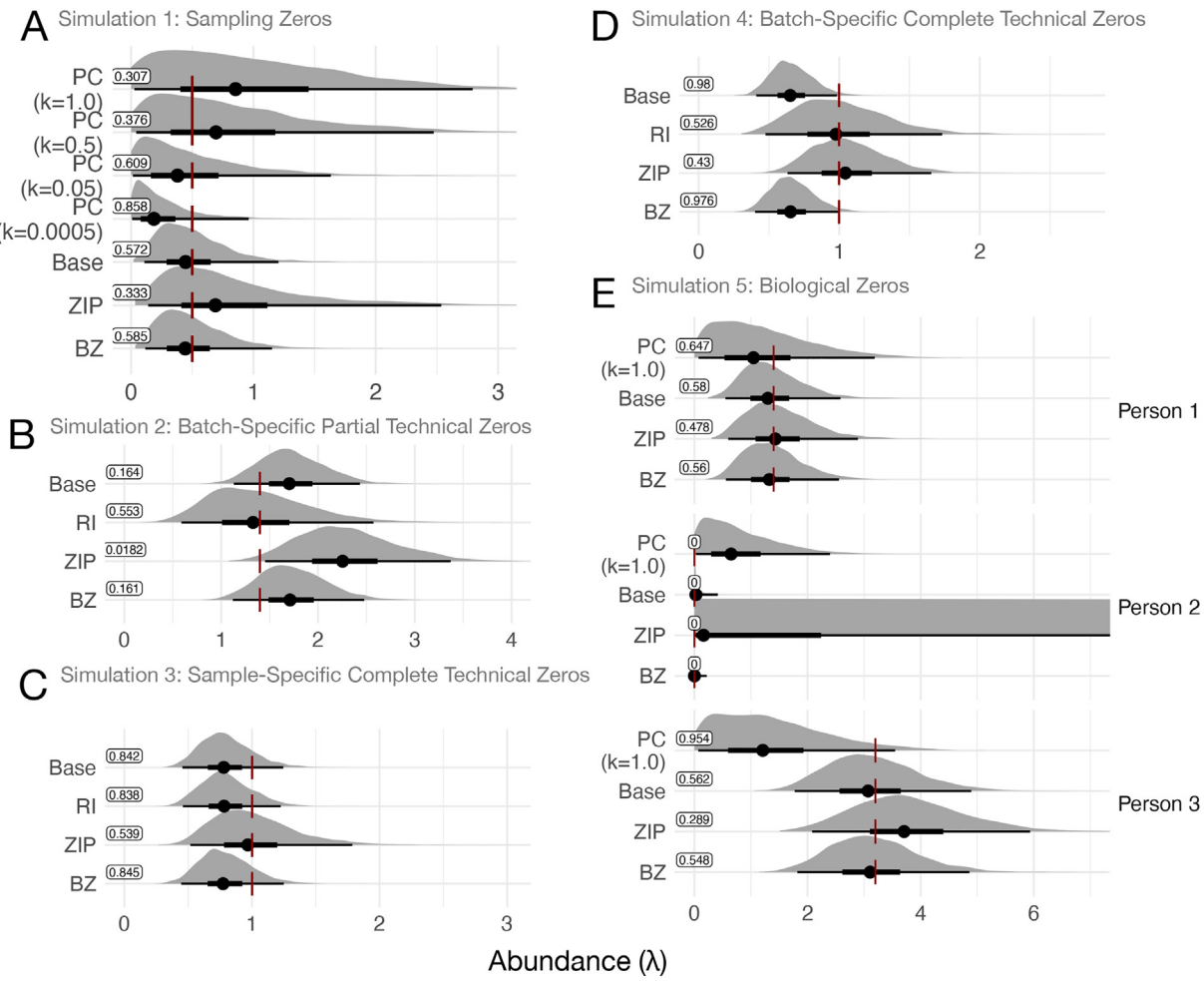

Fig. 3. A summary of how different zero handling models behave on different simulations of zero generating processes. Shown are posterior distributions of sequence abundance (λ) from each model. The dark red vertical bar represents true value of λ. Posterior mean as well as the 66% and 95% credible intervals are shown in black. Boxed values represent the cumulative distribution function of the true value of λ, as described in the text and in Figure S8, the best performance possible is a value of 0.5. The statistic is by definition zero when $\lambda^{true} = 0$ (*e.g.*, person 2 in panel E) in which case performance is assessed visually. (A) Simulation 1 (sampling zeros only), (B) Simulation 2 (batch-specific partial technical and sampling zeros), (C) Simulation 3 (sample-specific complete technical and sampling zeros) (D) Simulation 4 (batch-specific complete technical and sampling zeros), (E) Simulation 5 (biological and sampling zeros). In panel E, the abundance axis (λ) was cropped to enable all model results to be shown.

abundance sequence (Figure S10). Such uncertainty biased parameter estimates even when more than 1,000 replicate samples were available to the model (Figure S11). That is, even when the ZIP model had access to many replicate samples, the model falsely inferred that zero-inflation was present even when it was not. Our simulation of biological zeros also showed that, when used to model condition-specific zero-inflation like in the ZINB model of Section 2, posterior estimates of λ under the ZIP model can be the highest in conditions when true sequence abundances are the lowest (Person 2; Fig. 3E). More broadly, our simulations recapitulate our earlier findings involving published datasets and provide examples for how zero-inflated models can spuriously infer sequences to be present even when unobserved.

5. Discussion

Here we have demonstrated, using real-world datasets, that different methods for modeling zeros can lead to disagreement among almost half of sequences when carrying out differential expression analysis. We also categorized zero generating processes (ZGPs) and summarized evidence in favor of each. Last, we used simulations to explore how different zero handling methods perform when different ZGPs are present. In concert, these analyses suggest caution when considering zero-inflated models for handling zero values. These models may increase false negative rates in differential expression analyses, capture ZGPs that may actually be described by simpler biological processes, and are sensitive to model mis-specification.

Our conceptual framework and simulations provide insight into the mechanisms underlying the outcomes of empirical benchmarking studies [8,9,68,42,69]. Thorsen et al. [8] report that the zero-inflated Gaussian model metagenomeSeq increasingly biases estimates of differential abundance as data sparsity increases. Similarly, Dal Molin et al. [69] observe that models using zero-inflation, such as SCDE [20], or Monocle [70], have a higher false-negative rate for differential expression than alternative non-zero-inflated models such as DESeq [11] or edgeR [49]. Our results suggest these errors arise when zero-inflated models are applied to datasets where sample-specific complete technical processes are insubstantial.

Beyond explaining empirical phenomena, our framework suggests three guidelines for modeling sequence count data. First, for designers of new sequence count models, biological zeros can be approximated as sampling zeros. This recommendation is logical as biological zeros can be considered sampling zeros from a sequence whose abundance is small. Moreover, our results demon-

strate that sampling models will correctly interpret zeros as evidence of low-abundance sequence. Treating biological zeros as evidence of very low abundance sequence should encourage simplicity in future models, and is also consistent with the design of several existing tools such as DEseq2 [11], Aldex2 [71], MALLARD [13], Fido [72], GPMicrobiome [12] and MIMIX [14].

Our second guideline is that zeroinflated models should be avoided. Our re-analysis of datasets from single cell RNA-seq, bulk RNA-seq, and 16S rRNA microbiome sequencing experiments suggested that zero-inflated models can result in the spurious conclusion that sequences are not differentially expressed when clear presence-absence patterns exist between experimental groups. Additionally, our literature review and categorization of ZGPs revealed that common motivating phenomena for zero-inflated models such as dropout can actually be considered forms of other common ZGPs. Last, we found that when sample-specific complete technical processes are not present in data, zero-inflated models produce biased estimates. Overall, this guideline is aligned with recent research demonstrating that after controlling for biological zeros in droplet single-cell RNA-seq experiments, zero-inflation is not necessary to describe the zero patterns observed in sequencing data [18]. Rather, zeros in these experiments were nearly perfectly captured by negative binomial models lacking zero-inflation (*i.e.*, without the need to model complete technical zeros). Thus, when sequence count data have more zeros than can be adequately modeled by a Poisson distribution, our guideline suggests that these "excess-zeros" be modeled using sampling, partial technical, and biological processes. Still, were sample-specific complete technical zeros believed to be present in a given dataset, the use of zero-inflated models would be warranted.

Our third guideline for modeling zero values is to employ simple count models such as the Poisson, negative binomial, or multinomial that can incorporate technical covariates (e.g., batch). We underscore that this guideline does not require knowing the true ZGPs present in a study. Moreover, we do not expect that models will be able to reliably make this distinction either. Rather, our simulations show that simple models capable of accounting for both sampling and partial technical zeros can produce accurate inferences under a range of different ZGPs. Fortunately, such models have also already been implemented for a range of applications including generalized linear regression [72,14], non-linear regression [72,73], clustering [74], time-series analysis [13,12,72], and classification [75,76].

Our guidelines for zero handling will eventually need to be incorporated into broader pipelines for the analysis of sequence count data. Other outstanding challenges exist in this arena. Tasks like gene and sequence variant calling or data processing can be considered to be independent of zero handling as they are often done as an isolated step prior to data modeling. Recent advances in understanding the impact of these independent tasks [77–79] should therefore combine in a straightforward manner with our zero handling guidelines. On the other hand, some tasks in sequence count data analysis, such as data normalization, in sequence count data analysis are likely to require solutions that interact with a given zero handling framework. Data normalization choices, for example may convert integer counts into continuous variables [67] and thereby remove key information needed to understand sampling zeros. Furthermore, the goal of data modeling may further impact the choice of zero–handling method; tasks such as cluster analysis or differential expression may have different sensitivities to the choice of zero–handling. Future zero handling studies may therefore be combined with empirical studies of full sequence data analysis pipelines to provide additional insight into how modeling choices ultimately affect sequence count data analysis.

Even within the study of zero–handling, key questions remain. For example, in Section 2, we focused on how the presence or absence of zero-inflation impacted inferred differential expression effect sizes. We focused on effect sizes, as opposed to p-values or false discovery rates, as we felt that these provided a more interpretable means of understanding the origin of the differing results from the ZINB and NB models. Still, in practice, many researchers may be interested in hypothesis testing and in these cases a more detailed exploration of how different zero–handling methods impact false negative and false positive rates would be impactful.

6. Methods

6.1. Analysis of previously published data

We analyzed six previously published datasets. Each dataset was pre-processed such that only sequences observed in at least 3 samples with at least 3 counts were retained. Each dataset was normalized to the median sequencing depth of the dataset. For each dataset, the ZINB and NB models were fit using default parameter values, an intercept, and a binary variable denoting which of two groups samples belonged to. The NB model was created from the ZINB model by additionally specifying the matrix parameter O_pi. Large negative values for this parameter reduce zero inflation. The ZINB model used the default value for O_pi to allow zero inflation; the NB model set *O_pi* to be a matrix populated with the number -10^6 to ensure no zero inflation was used. The non-condition-specific ZINB model was created from the ZINB model by additionally specifying which_X_pi $= 1$ and which_V_pi $= 1$. Details regarding how each dataset can be obtained, the groups compared, and the resulting dataset sparsity level are given in Supplementary File 2.

Differential zero-inflation was reported directly by ZINB-WaVE. The ZINB-WaVE model infers a zero-inflated parameter π_j on the logit scale. For the condition-specific ZINB model, π_j is described by the linear model $\pi_j = \beta_{1j} + \beta_{2j}x_i$ where x_i is condition of sample x_i. Therefore, the paraeter β_{2j} can be interpreted as the degree to which the model differentially uses zero-inflation in one condition compared to another for sequence j – hence we termed this parameter differential zero-inflation.

6.2. Simulation studies

First we introduce our notation.

y_i	The number of counts of a specific sequence in the i^{th} sample
z_i	The biological sample where sample i originates
x_i	The batch in which sample i was processed

The following five models assume that each of the K biological specimens has a true parameter λ_k that represents the abundance of a single sequence j.

6.2.1. Prototypical models

For comparability, each of our five models are based on a hierarchical Poisson log-normal model. Letting κ denote a fixed non-zero value, we define the **pseudo-count (PC) model** as

$$(y_i + \kappa) \sim \text{LogNormal}(\lambda_{z_i}, \sigma^2)$$
$$\lambda_k \sim \text{Normal}(\rho, \tau^2).$$

This model avoids numerical issues with taking the log of zero values by adding a pseudo-count to the data prior to analysis.

We defined the **base model** as

$$y_i \sim \text{Poisson}(\lambda_{z_i})$$
$$\lambda_k \sim \text{LogNormal}(\mu, \sigma^2)$$
$$\mu \sim \text{Normal}(\rho, \tau^2).$$

This model considers count variation and zeros due to sampling.

The **random intercept (RI) model** modifies the base model with a batch-specific multiplicative factor η_{x_i}, which may alter the rate of Poisson sampling.

$$y_i \sim \text{Poisson}\left(\lambda_{z_i}\eta_{x_i}\right)$$
$$\eta_m \sim \text{LogNormal}(v, \omega^2)$$
$$\lambda_k \sim \text{LogNormal}(\mu, \sigma^2)$$
$$\mu \sim \text{Normal}(\rho, \tau^2).$$

For identifiability, we assume that a single batch, labeled batch number 1, is an unbiased gold standard, so $\eta_1 = 1$. If we have a single batch, this model is identical to the base model. The NB (negative binomial) model of Section 2 is similar to the RI model but uses the more flexible negative binomial distribution instead of the Poisson.

Like the ZINB model in Section 1, we created a **zero-inflated Poisson (ZIP) model** by adding a zero-inflated component to the Poisson part of the base model. The ZIP model is defined by

$$\begin{aligned} y_i &\sim \text{ZIP}(\lambda_{z_i}, \theta_{x_i}) \\ \lambda_k &\sim \text{LogNormal}(\mu, \sigma^2) \\ \theta_m &\sim \text{Beta}(\alpha, \beta) \\ \mu &\sim \text{Normal}(\rho, \tau^2) \end{aligned} \tag{1}$$

where $\text{ZIP}(\lambda_{z_i}, \theta_{x_i})$ is shorthand for

$$y_i \sim \begin{cases} \delta_0 & \text{if } w_i = 0 \\ \text{Poisson}(\lambda_{z_i}) & \text{if } w_i = 1 \end{cases}$$
$$w_i \sim \text{Bernoulli}(\theta_{x_i})$$

and where δ_0 refers to the Dirac distribution centered at zero. This model assumes that all zeros arise due to a sampling process or a complete technical process.

In contrast to the ZIP model, the **Biological Zero (BZ) model** adds a zero-inflated component to the log-normal part of the base model. The BZ model is defined by

$$y_i \sim \text{Poisson}(\lambda_{z_i})$$
$$\lambda_{z_i} \sim \text{ZILN}\left(\mu, \sigma^2, \gamma_{z_i}\right)$$
$$\gamma_k \sim \text{Beta}(\zeta, \xi)$$
$$\mu \sim \text{Normal}(\rho, \tau^2).$$

Here $\text{ZILN}(\mu, \sigma^2, \gamma_{z_i})$ is short for:

$$\lambda_i \sim \begin{cases} \delta_0 & \text{if } w_i = 0 \\ \text{LogNormal}(\mu, \sigma^2) & \text{if } w_i = 1 \end{cases}$$
$$w_i \sim \text{Bernoulli}\left(\gamma_{z_i}\right).$$

this model assumes that zeros arise from a sampling process or a biological process. In Section 4, the BZ model was modified from this form because of the difficulty of representing the latent Dirac distribution using the Hamiltonian Monte Carlo. Instead, the Dirac distribution in the BZ model was approximated with a truncated normal distribution with mean 0 and variance 0.0001.

6.2.2. Simulations

Our series of simulation studies investigated the behavior of each model on each zero generating process. We present only univariate simulations. Hyper-parameter values were chosen to use in each of the five simulations. The hyper-parameters are: $\sigma^2 = 3$, $\rho = -1$, $\tau^2 = 5$, $v = 0$, $\omega^2 = 2$, $\alpha = .5$, $\beta = .5$, $\zeta = 1$, and $\xi = 1$. Simulations that had low likelihood under the simulating model were rerun. This procedure ensured that each simulated dataset contained enough information to recover the true parameter values. This was done for simulations in Fig. 3 but not for simulations in Fig. S9.

Simulation 1: Sampling Zeros. The first simulation consisted of 5 random draws from a Poisson distribution with a rate parameter λ of 0.5. This represents a single sequence within a single person, measured with 5 technical replicates all processed in the same batch. We applied the PC model with three different pseudo-counts: 1,.5, and.05.

Simulation 2: Sampling and Batch-Specific Partial Technical Zeros. The second simulation consisted of 15 replicates samples split into 3 batches with Poisson rate parameters 1.4, 0.6, and 3.2. This simulates polymerized chain reaction (PCR) efficiency varying by batch. As discussed above, batch 1 is derived from some gold standard measurement device with no bias.

Simulation 3: Sampling and Sample-Specific Complete Technical Zeros. The third simulation consisted of 15 replicate samples from a Poisson distribution with rate parameter λ of 1. This simulates a single sequence measured with technical replicates where each replicate has a 30% chance of catastrophic error, causing a complete inability to measure that sequence. This simulation was contrived to reflect the assumptions of the ZIP model.

Simulation 4: Sampling and Batch-Specific Complete Technical Zeros This simulation uses sampling and complete technical zeros, and represents a single sequence measured in 15 replicate samples in 3 batches. However, because of a different reagent or missed experimental step, batch 2 complete lacked the sequence. We assume that no other bias is present in batches 1 or 3, which are represented as random draws from a Poisson distribution with rate parameter 1.

Simulation 5: Sampling and Biological Zeros The fifth simulation consisted of 15 samples from three individuals with Poisson rate parameters 1.4, 0, and 3.2. This simulates the abundance of a single sequence measured in three individuals, of which two possess that sequence and one does not. To model biological zeros with zero inflation, we slightly modify Eq. (1) in the ZIP model, replacing θ_{x_i} with θ_{z_i}. This change reflects a change of modeling zero-inflation by batch to modeling zero-inflation by individual. This corresponds to modeling condition-specific zero-inflation as in the ZINB-WaVE model.

6.2.3. Posterior inference

All 5 models were implemented in the Stan modeling language which uses Hamiltonian Monte Carlo (HMC) sampling [80]. Model inference was performed using 4 parallel chains, each with 1,000 transitions for warmup and adaptation and 1,000 iterations collected as posterior samples. Convergence of chains was determined by manual inspection of sampler trace plots and through inspection of the split $\widehat{R}$ statistic.

Code availability

All code necessary to recreate the analysis and figures in this work is available at:https://github.com/jsilve24/zero_types_paper.

CRediT authorship contribution statement

Justin D. Silverman: Conceptualization, Methodology, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Kimberly Roche:** Writing - review & editing. **Sayan Mukherjee:** Conceptualization, Writing - original draft, Supervision. **Lawrence A. David:** Conceptualization, Writing - review & editing, Supervision, Funding acquisition.

Acknowledgements

We thank Rachel Silverman for her manuscript comments. JDS and LAD were supported in part by the Duke University Medical Scientist Training Program (GM007171), the Global Probiotics Council, a Searle Scholars Award, the Hartwell Foundation, an Alfred P. Sloan Research Fellowship, the Translational Research Institute through Cooperative Agreement NNX16AO69A, the Damon Runyon Cancer Research Foundation, the Hartwell Foundation, and NIH 1R01DK116187-01. SM and KR would like to acknowledge the support of grants NSF IIS-1546331, NSF DMS-1418261, NSF IIS-1320357, NSF DMS-1045153, and NSF DMS1613261.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.csbj.2020.09.014.

References

[1] Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol 2016;17(1):75.
[2] Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 2017;5(1):27.
[3] Kaul A, Mandal S, Davidov O, Peddada SD. Analysis of microbiome data in the presence of excess zeros. Front Microbiol 2017;8:2114.
[4] Zhu L, Lei J, Devlin B, Roeder K. A unified statistical framework for single cell and bulk RNA sequencing data. Annals Appl Stat 2018;12(1):609.
[5] Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. eLife 2017;6:e21887.
[6] Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annu Rev Stat Its Appl 2015;2(1):73–94.
[7] Gloor GB, Macklaim JM, Vu M, Fernandes AD. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. Austrian J Stat 2016;45(4):73.
[8] Thorsen J, Brejnrod A, Mortensen M, Rasmussen MA, Stokholm J, Al-Soud WA, Sørensen S, Bisgaard H, Waage J. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. Microbiome 2016;4(1):62.
[9] Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nature Methods 2018;15(4):255.
[10] Aitchison J. The statistical analysis of compositional data. Monographs on statistics and applied probability. London; New York: Chapman and Hall; 1986.
[11] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
[12] Aijö T, Mü Ller CL, Bonneau R. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. Bioinformatics 2017.
[13] Silverman JD, Durand HK, Bloom RJ, Mukherjee S, David LA. Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. Microbiome 2018;6(1):202.
[14] Grantham NS, Reich BJ, Borer ET, Gross K. MIMIX: a Bayesian Mixed-Effects Model for Microbiome Data from Designed Experiments. arXiv, 2017..
[15] La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, Sodergren E, Weinstock G, Shannon WD. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLOS ONE 2012;7(12):1–13.
[16] Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nature Commun 2018;9(1):284.
[17] Wang J, Huang M, Torre E, Dueck H, Shaffer S, Murray J, Raj A, Li M, Zhang NR. Gene expression distribution deconvolution in single-cell RNA sequencing. Proc Nat Acad Sci 2018;115(28):E6437–46.
[18] Svensson V. Droplet scRNA-seq is not zero-inflated, bioRxiv; 2019, p. 582064..
[19] Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single cell RNA-seq based on a multinomial model. bioRxiv; 2019, p. 574574..
[20] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nature Methods 2014;11(7):740.
[21] Andrews TS, Hemberg M. False signals induced by single-cell imputation. F1000Research 2018;7.
[22] Gong W, Kwak I-Y, Pota P, Koyano-Nakagawa N, Garry DJ. DrImpute: imputing dropout events in single cell rna sequencing data. BMC Bioinform 2018;19(1):220.
[23] Leote AC, Wu X, Beyer A. Network-based imputation of dropouts in single-cell RNA sequencing data. bioRxiv; 2019..
[24] Lin P, Troup M, Ho JW. CIDR: ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol 2017;18(1):59.
[25] Pierson E, Yau C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol 2015;16(1):241.
[26] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 2015;16(3):133.
[27] Van den Berge K, Perraudeau F, Soneson C, Love MI, Risso D, Vert J-P, Robinson MD, Dudoit S, Clement L. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biol 2018;19(1):24.
[28] Ye C, Speed TP, Salim A. DECENT: differential expression with capture efficiency adjustment for single-cell RNA-seq data. Bioinformatics 2019;6.
[29] Alam M, Al Mahi N, Begum M. Zero-inflated models for RNA-Seq count data. J Biomed Anal 2018;1(2).
[30] Choi H, Gim J, Won S, Kim YJ, Kwon S, Park C. Network analysis for count data with excess zeros. BMC Genetics 2017;18(1):93.
[31] Oh S, Song S. Bayesian modeling approaches for temporal dynamics in RNA-seq data. New Insights into Bayesian Inference 2018:7.
[32] Zhou Y, Wan X, Zhang B, Tong T. Classifying next-generation sequencing data using a zero-inflated poisson model. Bioinformatics 2017;34(8):1329–35.
[33] Chen EZ, Li H. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. Bioinformatics 2016;32(17):2611–7.
[34] Chen L, Reeve J, Zhang L, Huang S, Wang X, Chen J. GMPR: a robust normalization method for zero-inflated count data with application to microbiome sequencing data. PeerJ 2018;6:e4600.
[35] Ho NT, Li F, Wang S, Kuhn L. metamicrobiomeR: an R package for analysis of microbiome relative abundance data using zero-inflated beta GAMLSS and meta-analysis across studies using random effects models. BMC Bioinform 2019;20(1):188.
[36] Jonsson V, Österlund T, Nerman O, Kristiansson E. Modelling of zero-inflation improves inference of metagenomic gene count data. Statistical Methods in Medical Research 2018. p. 0962280218811354.
[37] Lee KH, Coull BA, Moscicki A-B, Paster BJ, Starr JR. Bayesian variable selection for multivariate zero-inflated models: Application to microbiome count data. Biostatistics 2018;12.
[38] Li Q, Jiang S, Koh AY, Xiao G, Zhan X. Bayesian Modeling of Microbiome Data for Differential Abundance Analysis. arXiv e-prints, p. arXiv:1902.08741; 2019..
[39] Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nature Methods 2013;10(12):1200–2.
[40] Peng X, Li G, Liu Z. Zero-inflated beta regression for differential abundance analysis with metagenomics data. J Comput Biol 2016;23(2):102–10.
[41] Xia Y, Sun J, Chen D-G. "Modeling zero-inflated microbiome data," in Statistical Analysis of Microbiome Data with R. Springer; 2018. p. 453–96.
[42] Xu L, Paterson AD, Turpin W, Xu W. Assessment and selection of competing models for zero-inflated microbiome data. PloS ONE 2015;10(7):e0129606.
[43] Pollen AA, Nowakowski TJ, Shuga J, Wang X, Leyrat AA, Lui JH, Li N, Szpankowski L, Fowler B, Chen P, Ramalingam N, Sun G, Thu M, Norris M, Lebofsky R, Toppani D, Kemp DW, Wong M, Clerkson B, Jones BN, Wu S, Knutsson L, Alvarado B, Wang J, Weaver LS, May AP, Jones RC, Unger MA, Kriegstein AR, West JAA. Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat Biotechnol 2014;32(10):1053–8.
[44] Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nature Commun 2017;8:14049.
[45] Haglund F, Ma R, Huss M, Sulaiman L, Lu M, Nilsson I-L, Höög A, Juhlin CC, Hartman J, Larsson C. Evidence of a functional estrogen receptor in parathyroid adenomas. J Clin Endocrinol Metab 2012;97(12):4631–9.
[46] McMurrough TA, Dickson RJ, Thibert SM, Gloor GB, Edgell DR. Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. Proc Nat Acad Sci 2014;111(23):E2376–83.
[47] Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, et al. Genomic analysis identifies association of fusobacterium with colorectal carcinoma. Genome Res 2012;22(2):292–8.
[48] Gevers D, Kugathasan S, Denson LA, Vázquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, et al. The treatment-naive microbiome in new-onset crohn's disease. Cell Host Microbe 2014;15(3):382–92.
[49] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26(1):139–40.
[50] Xue JY, Zhao Y, Aronowitz J, Mai TT, Vides A, Qeriqi B, Kim D, Li C, de Stanchina E, Mazutis L, et al. Rapid non-uniform adaptation to conformation-specific kras (g12c) inhibition. Nature 2020;577(7790):421–5.

[51] Forsyth A, Raslan K, Lyashenko C, Bona S, Snow M, Khor B, Herrman E, Ortiz S, Choi D, Maier T, et al. Children with autism spectrum disorder: Pilot studies examining the salivary microbiome and implications for gut metabolism and social behavior. Human Microbiome J 2020;15:100066.
[52] Sa JM, Cannon MV, Caleon RL, Wellems TE, Serre D. Single-cell transcription analysis of plasmodium vivax blood-stage parasites identifies stage-and species-specific profiles of expression. PLoS Biol 2020;18(5):e3000711.
[53] Zerti D, Collin J, Queen R, Cockell SJ, Lako M. Understanding the complexity of retina and pluripotent stem cell derived retinal organoids with single cell rna sequencing: current progress, remaining challenges and future prospective. Curr Eye Res 2020;45(3):385–96.
[54] Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 2012;40. pp. e72–e72.
[55] Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 2011;12(2):R18.
[56] Liu CM, Price LB, Hungate BA, Abraham AG, Larsen LA, Christensen K, Stegger M, Skov R, Andersen PS, et al. Staphylococcus aureus and the ecology of the nasal microbiome. Sci Adv 2015;1(5).
[57] Shen S-Q, Yan X-W, Li P-T, Ji X-H. Analysis of differential gene expression by RNA-seq data in abcg1 knockout mice. Gene 2019;689:24–33.
[58] Farris M, Olson J. Detection of actinobacteria cultivated from environmental samples reveals bias in universal primers. Lett Appl Microbiol 2007;45 (4):376–81.
[59] Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010;11(10):733–9.
[60] Sinha R, Abu-Ali G, Vogtmann E, Fodor AA, Ren B, Amir A, Schwager E, Crabtree J, Ma S, et al. Assessment of variation in microbial community amplicon sequencing by the microbiome quality control (MBQC) project consortium. Nat Biotechnol 2017;35(11):1077–86.
[61] Polz MF, Cavanaugh CM. Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol 1998;64(10):3724–30.
[62] Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl Environ Microbiol 2005;71(12):8966–9.
[63] Silverman JD, Bloom RJ, Jiang S, Durand HK, Mukherjee S, David LA. Measuring and mitigating PCR bias in microbiome data. bioRxiv 2019:604025.
[64] Pinto AJ, Raskin L. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. PloS ONE 2012;7(8).
[65] Wear EK, Wilbanks EG, Nelson CE, Carlson CA. Primer selection impacts specific population abundances but not community dynamics in a monthly time-series 16S rRNA gene amplicon analysis of coastal marine bacterioplankton. Environ Microbiol 2018;20(8):2709–26.
[66] Eddy SR. What is Bayesian statistics?. Nature Biotechnol 2004;22(9):1177–8.
[67] McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol 2014;10(4).
[68] Quinn TP, Crowley TM, Richardson MF. Benchmarking differential expression analysis tools for RNA-seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinform 2018;19(1):274.
[69] Dal Molin A, Baruzzo G, Di Camillo B. Single-cell RNA-sequencing: assessment of differential expression analysis methods. Front Genet 2017;8:62.
[70] Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nature Methods 2017;14(3):309.
[71] Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2014;2:15.
[72] Silverman JD, Roche K, Holmes ZC, David LA, Mukherjee S. Bayesian multinomial logistic normal models through marginally latent matrix-T processes. arXiv e-prints, p. arXiv:1903.11695; 2019.
[73] Ren X, Kuan PF. Negative binomial additive model for RNA-Seq data analysis. bioRxiv; 2019..
[74] Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLOS ONE 2012;7:1–15.
[75] Gao X, Lin H, Dong Q. A Dirichlet-multinomial Bayes classifier for disease diagnosis with microbial compositions. mSphere 2017;2(6).
[76] Dong K, Zhao H, Tong T, Wan X. NBLDA: negative binomial linear discriminant analysis for RNA-seq data. BMC Bioinform Sep 2016;17:369.
[77] Baruzzo G, Hayer KE, Kim EJ, Di Camillo B, FitzGerald GA, Grant GR. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nature Methods 2017;14(2):135.
[78] Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. Dada2: High-resolution sample inference from illumina amplicon data. Nat Methods 2016;13(7):581–3.
[79] Everaert C, Luypaert M, Maag JL, Cheng QX, Dinger ME, Hellemans J, Mestdagh P. Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. Sci Rep 2017;7(1):1559.
[80] Gelman A, Lee D, Guo J. Stan: A probabilistic programming language for Bayesian inference and optimization. J Educ Behav Stat 2015;40(5):530–43.