Naught all zeros in sequence count data are the same

Justin D. Silverman[1,2], Kimberly Roche[1], Sayan Mukherjee[1,3,4,6], and Lawrence A. David[1,4,5,6]

[1] *Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27708*
[2] *Medical Scientist Training Program, Duke University, Durham, NC 27708*
[3] *Departments of Statistical Science, Mathematics, Computer Science, Biostatistics & Bioinformatics, Duke University, Durham, NC 27708*
[4] *Center for Genomic and Computational Biology, Duke University, Durham, NC 27708*
[5] *Department of Molecular Genetics and Microbiology, Duke University, Durham, NC 27708*
[6] *Denotes Co-corresponding Authors*

Abstract

Due to the advent and utility of high-throughput sequencing, modern biomedical research abounds with multivariate count data. Yet such sequence count data is often extremely sparse; that is, much of the data is zero values. Such zero values are well known to cause problems for statistical analyses. In this work we provide a systematic description of different processes that can give rise to zero values as well as the types of methods for addressing zeros in sequence count studies. Importantly, we systematically review how various models perform on each type of zero generating process. Our results demonstrate that zero-inflated models can have substantial biases in both simulated and real data settings. Additionally, we find that zeros due to biological absences can, for many applications, be approximated as originating from under sampling. Beyond these results, this work provides a paired categorization scheme for models and zero generating processes to facilitate discussions and future research into the analysis of sequence count data.

1 Introduction

Sequence counting refers to the use of high-throughput DNA sequencing to profile the abundance of distinct DNA or RNA transcripts within a biological system. Such sequence counting is widely used in biomedical research [1] including the investigation of host-associated microbial communities (e.g., 16S rRNA sequencing) [2], gene expression in single (scRNA-seq) or groups of cells (RNA-seq) [3, 4], lymphocyte populations [5], or even the interaction of nucleotide binding molecules with DNA or RNA [6, 7]. Zero counts are highly prevalent in many high-throughput DNA sequencing assays. For example, it is not uncommon to see microbiome or single cell RNA-seq data containing over 70% zero values [4, 8, 9]. Importantly, zero values can pose problems for available modeling approaches [10, 9, 11, 12]. Beyond the challenges of taking the log or dividing by zero, modeling in the presence of zero values is challenging because there are multiple sources that generate zero values . For example, zero values may arise due to abundances below the limits of detection, technical biases such as batch effects, or complete absence of a transcript from the biological system of interest [4, 13, 9].

While many models have been developed for the analysis of sequence count data, there remains a gap in the current literature regarding how the assumptions made by different models effect inference in the presence of zero values. Most notably, there have been no systematic investigations of the impact of deviations from modeling assumptions on results with respect to different zero generating processes. To address this current limitation, this work introduces a paired classification system for zero generating processes and classes of models with respect to their handling of zero values. By framing this classification system in terms of the processes that can generate zero values, we are able to provide examples of each type of zero from a biological perspective. Most importantly, this work elucidates the impact on inference of using each type of model in the presence of each type of zero generating process. We demonstrate how mismatches between models and zero generating processes can lead to spurious conclusions and provide guidance for how to avoid such mismatches.

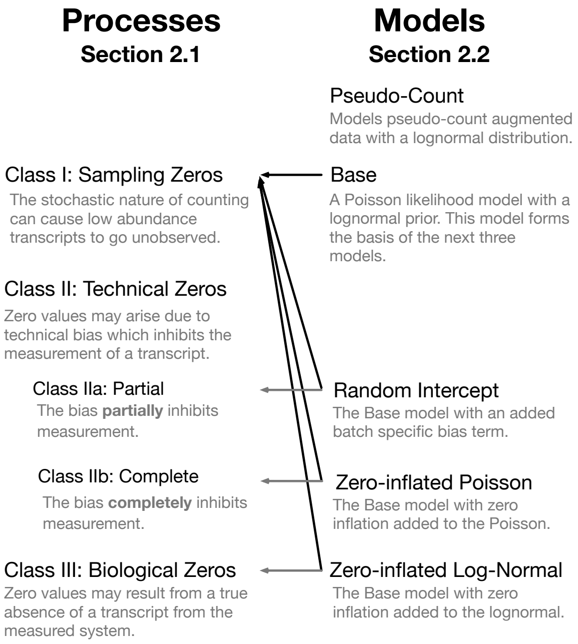

Processes
Section 2.1
Models
Section 2.2
Pseudo-Count
Models pseudo-count augmented data with a lognormal distribution.
Class I: Sampling Zeros
The stochastic nature of counting can cause low abundance transcripts to go unobserved.
Base
A Poisson likelihood model with a lognormal prior. This model forms the basis of the next three models.
Class II: Technical Zeros
Zero values may arise due to technical bias which inhibits the measurement of a transcript.
Class IIa: Partial
The bias partially inhibits measurement.
Random Intercept
The Base model with an added batch specific bias term.
Class IIb: Complete
The bias completely inhibits measurement.
Zero-inflated Poisson
The Base model with zero inflation added to the Poisson.
Class III: Biological Zeros
Zero values may result from a true absence of a transcript from the measured system.
Zero-inflated Log-Normal
The Base model with zero inflation added to the lognormal.

Figure 1: An overview of the zero generating processes and models presented in this work. The zero generating processes that each model accounts for are depicted with arrows.

To provide a framework for contextualizing zero values in sequence count data, we introduce a classification scheme for zero values based on the generative processes that can introduce zeros into data. This approach, classifying zero based on generative processes, is an extension of the commonly discussed distinction between sampling zeros and structural zeros (*i.e.,* rounded and essential zeros) [14]. However, the distinction of sampling versus structural zeros is not refined enough to describe the assumptions underlying many currently available models (e.g., zero-inflated models). To extend this classification scheme, we define three classes of zeros: sampling zeros, technical zeros, and biological zeros (Figure 1). Beyond providing a framework for discussing zero generating processes, this classification scheme is designed to pair with a simple classification of existing models.

Zero counts are often well mixed in sequence count data and cannot be easily removed by filtering select transcripts or samples, thus a number of techniques have been developed for modeling zeros in sequence count studies. Here we classify the most common methods for handing zero values into four types. The simplest approach is to add a small positive number (*e.g.*, 0.5 or 1), called a pseudo-count, to all observed counts and then perform analyses on the resulting pseudo-count augmented data (class 0 models) [9]. Other approaches rely on modeling zero values as resulting from stochastic processes (class I-III models). Class I models use a variety of distributions such as the Poisson [15], or negative binomial [3, 16] to model stochastic sampling. Class II models go beyond modeling stochastic sampling by considering that zero values may arise due to secondary sources. Class IIa models, such as the multinomial-Dirichlet [17, 11, 12] or multinomial-logistic normal [18, 19, 20] based models represent sources of technical variation beyond stochastic sampling that can introduce zeros into the data. In contrast, class IIb models such as zero inflated models [18, 21, 22, 23] or hurdle models [24, 13], consider secondary sources of technical variation or bias that specifically add zero values into observed data. Finally, class III models refer to those models that consider zeros to potentially originate from true biological absences in addition to stochastic sampling. While multiple methods are intended to model biological absences (notably [9] and [25]) these models instead exclude biological zeros from computations using either ad-hoc methods or marginalization.

To systematically investigate the behavior of each model class on each type of zero generating processes, we made use of both simulated and real data. We find that simpler sampling models (e.g., class I models) are well equipped to perform inference in situations dominated by sampling and biological zeros. In contrast, we find that zero-inflated models (an example of class IIa models) tend to inflate parameter estimates in both simulated and real data settings due to inherent identifiability[1] issues with such models. In fact, this parameter inflation can be so severe as to dominate the results of a differential expression analysis on a previously published single-cell RNA-seq study. Taken together our results and discussion present a framework for contextualizing zero values in sequence count studies and suggest caution in the use of zero inflated models.

2 Zero Processes and Prototypical Analysis Methods

An overall goal of this work is to describe the interplay between processes that can give rise to zero values and methods used for modeling in the presence of zero values. In service to this goal, we provide a categorization scheme for zero generating processes in Section 2.1. In Section 2.2, we introduce a categorization of zero modeling methods designed to parallel our characterization of zero generating processes. Finally, in section 2.3 we extend this discussion to multivariate counting models such as the multinomial.

2.1 Zero Processes

Here we classify processes that can introduce zero values into sequence count studies. Our categorization scheme involves three classes of processes (sampling, technical, and biological), the second of which we further subdivide into partial technical and complete technical. In brief, this categorization can be seen to extend the common rounded versus essential/structural classification of zeros (*i.e.*, zeros that do versus do not disappear with greater counting effort)[14]. In particular, we identify rounded zeros with biological and partial technical zeros and essential zeros with complete technical and biological zeros. The remainder of this section is dedicated to defining and giving examples of each of these classes of zero generating process.

2.1.1 Class I: Sampling Zeros

Sequencing depth (the total number of transcripts counted in a given sample) is limited in that only a small percentage of the total number of transcripts in a biological system of interest end up being counted. Due to limited sequencing depth in combination with the stochastic nature of counting, transcripts with low but non-zero abundance will occasionally go undetected thus introducing zero values into sequence count data [9].

We consider sampling zeros to be the most fundamental type of zero in sequence count studies as the process of counting itself guarantees such zeros are possible. In contrast, technical and biological zero generating processes may only exist in real data if additional processes are present in addition to sampling processes. For this reason, all the simulations we describe in Section 4 will contain sampling zeros with some additionally containing secondary technical or biological zeros.

2.1.2 Class II: Technical Zeros

The experimental steps required to prepare a sample for sequencing can also introduce zero counts into data by partially (class IIa) or completely (class IIb) reducing the amount of a DNA transcript available to be counted. As an example of partial technical zeros: the bacterial phylum Actinobacteria may go unobserved due to a *relative* inability to lyse and extract DNA from these cells [26]. As a result, the abundance of Actinobacterial DNA in processed samples will be lower than comparable sampled processed without this bias. This can in turn lead to zero values in sequence count data.

If, instead of a *relative* inability to lyse Actinobacteria there was a *complete* inability to lyse Actinobacteria, we would refer to the process as a complete technical process. While the distinction between partial and complete technical zeros may seem minute, consider that partial technical zeros disappear with increasing sequencing depth whereas complete technical zeros do not. As discussed in Section 2.2, this difference in their behavior with increasing sequencing depth also leads to different models for handling partial versus complete technical zeros.

[1] A model is called **identifiable** if given data supports a single value for each parameter. In contrast, a model is non-identifiable if multiple parameter values lead to the same model fit.

2.1.3 Class III: Biological Zeros

Zero counts can also arise from true absence of a transcript in a biological system. For example, the absence of a gene within a cell or the absence of a bacterial species within an individual may both introduce zero values into sequence count data. In contrast to sampling or technical processes which reflect observed zeros from transcripts that have non-zero abundance, biological zeros represent a transcript that is truly absent from a biological system. Seen another way, biological zeros are most similar to sampling zeros in that they both represents a truly low relative abundance, not a relative abundance that is artificially deflated as in technical zeros. Additionally, like complete technical zeros, these zeros are present even with unlimited sequencing depth.

2.2 Five Prototypical Models

In this subsection we describe a categorization scheme of zero modeling methods designed to parallel our characterization of zero generating processes in Section 2.1. These models are not meant to reflect the full complexity of analyzing sequence count data but instead to serve as illustrative tools. For this reason, here we consider only univariate Bayesian models where a single entity is counted. In Section 2.3 we extend this discussion to multivariate models. Additionally, to maintain a coherency between the example models, all models presented are based on the hierarchical Poisson Log-Normal distribution. This hierarchical model is highly flexible and has similarity to the negative binomial distribution which has been used frequently in the analysis of sequence count data. Before describing the models we highlight several modeling principles to provide intuition for the behavior of these models as well as basic notation that will be used throughout this and subsequent sections.

The hierarchical models present can be thought of as having an ordering going from the likelihood (written at the top of a hierarchical model) to the prior (lower in a hierarchical model). This ordering is of crucial importance in intuiting modeling results. Whereas methods that introduce extra zeros (*e.g.*, zeros from technical or biological processes beyond sampling zeros) at the bottom of a model (as part of the prior) will act to deflate abundance estimates, methods that introduce extra zeros at the top of a model (as part of the likelihood) will paradoxically act to increase abundance estimates. This later feature is driven by the following logic: if an extra process is present that introduces zeros because of technical artifact, the true abundance must be higher than the observed data would suggest. In other words, a model that considers extra zero generating processes at the top implicitly remove some zero observations as technical noise thus estimating abundance based on the higher non-zero counts. This feature will present a recurring theme in a number of models, most notably zero-inflated models.

In the following sections we will make use of the following notation. Let Y represent sequence count data such that Y_{ij} represents the number of counts observed from biological entity $j \in \{1, ..., D\}$ in sample $i \in \{1, ..., N\}$. Additionally, define $z_i \in \{1, ..., K\}$ as the biological specimen (*e.g.*, the person) which sample i originates and $x_i \in \{1, ..., M\}$ as 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 transcript j.

2.2.1 Class 0: Fixed Zero Replacement Models

The first class of models we consider replaces zero values with a fixed non-zero value (*e.g.*, a pseudo-count) which we denote below as κ. As a prototype of this class, we consider a model in which the logarithm of the pseudo-count augmented data are distributed as a normal with a mean dependent on which biological specimen the sample came from. Here we place a normal prior on the K mean parameters.

$$(y_i + \kappa) \sim \text{LogNormal}(\lambda_{z_i}, \sigma^2) \tag{1}$$

$$\lambda_k \sim \text{Normal}(\rho, \tau^2) \tag{2}$$

We will refer to the above example of a class 0 model as the **pseudo-count (PC) model**.

The primary purpose of models of this class is to avoid numerical issues often introduced when taking the logarithm or diving by zeros. As we will demonstrate in Section 4, this class of models ignores the count variation present in the data and can be extremely sensitive to the choice of pseudo-count. In comparison to the other classes of models we will present, this class performs poorly under all zero generating processes.

2.2.2 Class I: Unadjusted Count Models

In this class we consider models based upon count distributions such as the Poisson or the negative binomial[2]. As a prototype of this class we consider a simple Poisson model in which each of the K biological specimens is modeled as having its own rate parameter λ which come from a shared log-normal distribution.

$$y_i \sim \text{Poisson}(\lambda_{z_i}) \tag{3}$$
$$\lambda_k \sim \text{LogNormal}(\mu, \sigma^2) \tag{4}$$
$$\mu \sim \text{Normal}(\rho, \tau^2) \tag{5}$$

Note that in this model, all zeros are introduced at the top of the model through the Poisson counting process. We will refer to the above example of a class I model as the **base model** as it will be the starting point for the remaining models we will introduce.

This class of models forms the foundation of many statistical methods for the analysis of sequence count data. Popular methods such DESeq2 [3] are most similar to this class. This class of models assumes that all zeros are sampling zeros (class I) and provides no mechanism to account for potentially higher rates of zero values due to additional technical or biological processes. Yet, as we will demonstrate in Section 4, this class of models performs well in the presence of sampling and biological zeros. In contrast, in the presence of technical zeros these models have a tendency to underestimate transcript abundance as this model provides no mechanism from excluding the influence of these zero values in parameter estimates.

2.2.3 Class IIa: Partial Technical Bias Models

In this class we consider models that account for technical bias as seen in a partial technical process. For example, consider a situation where DNA from a specified transcript is amplified with different efficiency between batches (*e.g.*, PCR bias differs between batches). For simplicity, here we consider this bias to be unknown but a single batch (which we label as batch number 1) is considered an unbiased gold standard. In this situation a prototypical model of class IIa may appear as follows:

$$y_i \sim \text{Poisson}(\lambda_{z_i}\eta_{x_i}) \tag{6}$$
$$\eta_m \sim \text{LogNormal}(\nu, \omega^2) \tag{7}$$
$$\lambda_k \sim \text{LogNormal}(\mu, \sigma^2) \tag{8}$$
$$\mu \sim \text{Normal}(\rho, \tau^2) \tag{9}$$

where η_m represents a multiplicative bias found in samples from batch m. To reflect our knowledge that batch number 1 is the closest to unbiased we also impose the constraint $\eta_1 = 1$. Note that in this model, all zero values are introduced at the top of the model through the Poisson counting process alone or through the Log-normal process on η_k which acts in concert with the Poisson process to introduce zero values. We refer to the above example of a class IIa model as the **random intercept (RI) model** as each batch is modeled as having its own random log-linear intercept term.

While we present this class as a method of modeling zero counts, it is more commonly considered as a method of modeling batch effects or other technical variation [27]. This model assumes that all zeros are sampling zeros but that the poison rate may be altered by technical factors (*e.g.*, PCR bias) which may in turn leads to higher rates of zero values. As we will demonstrate in Section 4, this class of models performs well in the presence of partial technical zeros yet adds little compared to the base model (class I) in the presence of other types of zeros. Overall, models of this class often require either strong prior knowledge regarding batch effects or technical replicate samples to identify the parameters η_m.

[2] The negative binomial distribution can be thought of as a Poisson distribution where the rate parameter λ is itself distributed according to the gamma distribution. While the resultant over-dispersion of the negative binomial can produce higher rates of zero values compared to the Poisson distribution, we still consider it a zero unadjusted count model as it does not explicitly model extra zero values as the random intercept, zero-inflated Poisson and zero inflated log-normal models do do. In practice, such over dispersion may produce estimates that lie in between those of class I and class III models.

2.2.4 Class IIb: Complete Technical Bias Models

In this class we consider models that assume a distinct second stochastic process which may introduce zero values into observed data. The most widely used models of this form are zero inflated models [18, 21, 22, 23] and hurdle models [24, 13]. We focus on a zero inflated Poisson (ZIP) model as a prototypical example of this type. In contrast to the base model (class I), this model says that zero counts may arise through a sampling process (the Poisson distribution) or a second zero inflation process (an example of a complete technical process). The zero inflation process has the following narrative: for every Poisson count, flip a coin that has a probability θ of coming up heads. If the coin shows a tail, keep the Poisson count; however, if the coin shows a heads, replace the Poisson count with zero. Here for simplicity, we model θ as depending only on the batch number and impose a beta prior on this quantity.

$$\begin{aligned} y_i &\sim \text{ZIP}(\lambda_{z_i}, \theta_{x_i}) && (10)\\ \lambda_k &\sim \text{LogNormal}(\mu, \sigma^2) && (11)\\ \theta_m &\sim \text{Beta}(\alpha, \beta) && (12)\\ \mu &\sim \text{Normal}(\rho, \tau^2) && (13)\end{aligned}$$

Here $\text{ZIP}(\lambda_{z_i}, \theta_{x_i})$ is shorthand for the following:

$$\begin{aligned} y_i &\sim \begin{cases} \delta_0 & \text{if } w_i = 0 \\ \text{Poisson}(\lambda_{z_i}) & \text{if } w_i = 1 \end{cases} && (14)\\ w_i &\sim \text{Bernoulli}(\theta_{x_i}) && (15)\end{aligned}$$

where δ_0 refers to the Dirac distribution which has positive density only at $y_i = 0$ and is zero otherwise[3]. Similar to the base and random intercept models, this model assumes that all zero counts are introduced at the top of the model (e.g., due to technical sources). We will refer to this model as the **Zero-Inflated Poisson (ZIP) model**.

This class of models has become increasingly popular and is widely used in tools such as ZINB-WaVE [21]. This model assumes some zeros are from sampling while other are due to a complete technical process. Notably, the presence of sampling zeros suggest that a transcript is at relatively low abundance; in contrast, the presence of complete technical zeros suggest that a transcript is potentially at high or low abundance. Yet, determining whether zeros come from a sampling or a complete technical process is often impossible and as such these models are weakly identified. That is, the model often cannot distinguish as to whether a zero is due to a high abundance transcript that was censored or a low abundance transcript that may or may not have been censored. As we will see in Sections 4 and 3 this uncertainty can lead to spurious conclusions. In particular, while the ZIP model performs well in the presence of select complete technical processes, it can substantially inflate inferred abundance in many other situations.

2.2.5 Class III: Biological Absence Models

In this class we consider models that, in addition to modeling sampling zeros model biological zeros as part of the prior distribution. Here, to draw analogy to the base and ZIP models, we utilize a zero-inflated log normal distribution (ZILN) [28]. We will model the parameters λ_k as either coming from a distribution with a point mass at zero or from a log-normal distribution as before. We introduce the parameter γ_k to represent the probability that $\lambda_k = 0$ analogously to θ_m used in the ZIP model.

$$\begin{aligned} y_i &\sim \text{Poisson}(\lambda_{z_i}) && (17)\\ \lambda_{z_i} &\sim \text{ZILN}(\mu, \sigma^2, \gamma_{z_i}) && (18)\\ \gamma_k &\sim \text{Beta}(\zeta, \xi) && (19)\\ \mu &\sim \text{Normal}(\rho, \tau^2). && (20)\end{aligned}$$

[3] Equivalently the probability mass function of the zero-inflated Poisson distribution can be written as

$$\text{P}(y \mid \lambda, \theta) = \begin{cases} \theta + (1-\theta)e^{-\lambda} & \text{if } y = 0 \\ (1-\theta)\frac{\lambda^y e^{-\lambda}}{y!} & \text{if } y > 0. \end{cases} \qquad (16)$$

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

$$\lambda_i \sim \begin{cases} \delta_0 & \text{if } w_i = 0 \\ \text{LogNormal}(\mu, \sigma^2) & \text{if } w_i = 1 \end{cases} \tag{21}$$

$$w_i \sim \text{Bernoulli}(\gamma_{z_i}). \tag{22}$$

We will refer to the above example of a class III model as the **Zero-Inflated Log-Normal (ZILN) model**.

While models of this type are infrequently discussed in terms of modeling zero values in sequence count data, they are common in many other areas of statistics. For example, such models are commonly used for variable selection and include the use of spike-and-slab, Laplace, and horseshoe prior distributions [29]. This model assumes that some zeros are due to sampling while other are due to biological absence. This later possibility, that zeros originate from a biological process, can lead to overall deflation of estimates compared to the base model. Yet this deflation is seen only rarely when nearly all counts are zero as even a single non-zero count makes it highly unlikely that $\lambda_k = 0$. As we demonstrate in Section 4, this class of models performs similarly to the base model in the presence of sampling, partial technical, and complete technical processes yet provides a marginal improvement over the base model in the presence of biological processes.

2.3 Multivariate Count Models

For the sake of interpretability, our discussion of zero values to this point has focused on zero generating processes and modeling approaches using univariate examples. Yet, sequence count data is fundamentally a multivariate measurement [12, 10]. For the most part, the conceptual tools introduced with respect to the univariate setting encompass much of the complexity of the multivariate setting. Here we extend our discussion to the multivariate setting by discussing a uniquely multivariate phenomena wherein zeros can be introduced in multivariate count data.

Zeros can be introduced into sequence count data through a competition-to-be-counted process which has technical origins and is a type of partial technical process. Biological samples go through many distinct steps, such as DNA extraction, PCR amplification, and library pooling, before they are ultimately sequenced. Each of these involves a random sampling where a random subset of DNA is sampled and carried over to subsequent processing steps. These random sampling events introduce a negative correlation between transcripts due to transcripts competing to be counted [30, 31]. That is, measuring more of one type of transcript decreases the chance that another transcript will be measured. This is therefore a type of partial technical process (Class IIa) as more of one transcript partially (not completely) inhibits our ability to measure another transcript. This competition to be counted has been successfully measured using multinomial based models [20, 18, 31]. As multinomial models model both this multivariate partial technical process (Class IIa) and sampling zeros (Class I) we consider multinomial models to be a hybrid of Class IIa and Class I models.

3 Real Data Example

To demonstrate that modeling different classes of zero values can lead to non-trivial differences in conclusions in real data scenarios, we reanalyzed a single cell RNA-seq study originally published by Pollen et al [32]. This data features measurements of gene expression from 301 single cells from 11 populations within the developing cerebral cortex. This data was subsequently reanalyzed to demonstrate the capabilities of ZINB-WaVE, a zero-inflated negative binomial linear model designed for the analysis of single cell RNA-seq data [21][4].We chose to reanalyze this study with ZINB-WaVE using the same data pre-processing and modeling steps used by Risso et al [21]. As in Risso et al [21] we analyze only the 100 most abundant genes in the dataset. We modify this procedure in only two ways so as to investigate differential abundance between the biological conditions NPC and GW21. First, we subset the sequenced cells to only contains cells from groups NPC and GW21. Second, we include a binary covariate for NPC versus GW21 in the ZINB-WaVE model to allow differential abundance to be estimated. As in the analysis of Risso et al [21], we

[4] In contrast, to the models discussed in Section 2.2 and 4 which made use of Bayesian posterior estimates, ZINB-WaVE uses penalized maximum likelihood for inference. This analysis serves the dual goal of demonstrating that the concepts we introduced in prior sections are not unique to the Bayesian models introduced in Section 2.2

also include a binary covariate for the coverage depth in the linear model to account for potential batch effects related to high versus low coverage.

The above model was applied to the data twice, once with and once without zero inflation to investigate the effects of the zero inflation component of the model (*Methods*). In the absence of zero-inflation this model, which we refer to as the NB model, considers zero values as coming only from a sampling process (the negative binomial component) and a partial technical process (the coverage depth predictor) processes. With the zero-inflation component this model, which we refer to as the ZINB model, considers zero values as coming from sampling, partial technical, as well as complete technical processes. Thus the NB model is most similar to the Base model presented in Section 2.2.2 whereas the ZINB model is most similar to the zero-inflated Poisson model presented in Section 2.2.4.

To identify inferences that differed between the NB and ZINB model, we looked at the 10 genes that had the largest difference in $\log_2$ fold-change differential expression ($\mathrm{DE}_{\mathrm{NB}} - \mathrm{DE}_{\mathrm{ZINB}}$) between the ZINB and NB models. Positive values of differential expression imply that the given gene is higher in the NPC condition whereas negative values reflect that the gene is higher in the GW21 condition. The resulting 10 genes, are shown in the top half of Table 1 and their distribution is shown in Figure S4. Notably, these 10 genes all contained a large number of zero values in one of the two conditions. For the gene SHISA2, all observations save a single count of 1 is zero, whereas only 1 value from the NPC condition is zero while the others are large and non-zero. Intuitively, such a scenario would suggest that SHISA2 is almost completely absent from the NPC condition but highly abundant in the GW21 condition (*e.g.*, SHISA2 is highly deferentially expressed). As expected, in the NB model, SHISA2 is estimated as the most highly expressed gene with a $\log_2$ differential expression value of 5.2. In contrast, in model ZINB, SHISA2 is estimated as the 28th most deferentially expressed gene with a $\log_2$ differential expression value of 0.86. While we cannot say which model is correct, it is clear that assumptions regarding which processes are introducing zero values into the data can cause large discrepancies in modeling results. Overall these results demonstrate that the inclusion of zero-inflation substantially alters modeling results.

To further identify differences between the ZINB and NB models we investigated the 10 genes found to have conflicting signs of differential expression between the two models. These 10 genes are shown in the bottom half of Table 1 and their distribution is shown in Figure S4. In contrast to the 10 genes with the largest difference in differential expression, all 10 of these genes had large non-zero counts in both NPC and GW21 conditions making it more plausible that these zeros are due to a class IIb process. For example, gene USP47 had a mean of 614 and 448 in the NPC and GW21 conditions respectively. This result includes zero values in the calculation of the mean and is therefore analogous to inference using the NB model. However, the ordering is reversed for gene USP47 when excluding zero values (means of 658 and 897 in the NPC and GW21 conditions respectively). This result excludes zero values in the calculation of the mean and is therefore analogous to inference using the ZINB model. Beyond further demonstrating that zero inflation can substantially change modeling results, these results suggest the largely counter intuitive behavior of such models. In changing the sign of differential abundance based on zero counts, zero inflated models are implicitly assuming that any zeros originating form a complete technical process are in fact as high abundance as the surrounding zero counts. Such behavior is extremely intuitive. In fact, we expect most researchers would expect and believe that zero counts represent either absent or low abundance transcripts, not transcripts that are high enough abundance to flip the signs of differential expression.

In order to further demonstrate how the concepts from Sections 2.1 and 2.2 may provide insight into the behavior of existing models on real data, we defined a quantity which we hypothesized would correlate well with large discrepancies between differential expression as measured by these two models. We refer to this quantity as the Zero Inclusion Ratio of Means which we define for each gene g as $\mathrm{ZIRM}_g = \log_2 \frac{\overline{x}_g^{\mathrm{NPC}}}{\overline{x}_g^{\mathrm{GW21}}} - \log_2 \frac{\overline{x}_{g/0}^{\mathrm{NPC}}}{\overline{x}_{g/0}^{\mathrm{GW21}}}$ where $\overline{x}_g^{\mathrm{NPC}}$ refers to the mean of the counts of gene g in the NPC condition, and $\overline{x}_{g/0}^{\mathrm{GW21}}$ refers to the mean of the counts of gene g in the GW21 condition excluding zero counts. The intuition behind this quantity is as follows: by considering that zero values can be from a class IIb process, zero-inflated models essentially exclude some zero values from contributing to the estimation of that gene's abundance, instead relying more heavily on the non-zero values. Thus we hypothesized that the log-ratio of the mean count value for a gene excluding zero values would be an approximation to the inferred differential expression of that gene from the ZINB model. Without this zero inflation component, the NB model uses all the zero

Table 1: Differential expression (DE) estimates from a negative binomial (NB) and zero-inflated negative binomial (ZINB) model can differ substantially. Differential expression was calculated in log base 2 with higher values representing a gene that is more abundant in the NPC as compared to GW21 condition. The first 10 genes represent the 10 genes that had the largest discrepancy between the differential expression estimated from the NB vs. ZINB models. The last 10 genes represent genes that were estimated as being differentially expressed but with opposing signs between the two models. Ranked differential expression is given in parentheses.

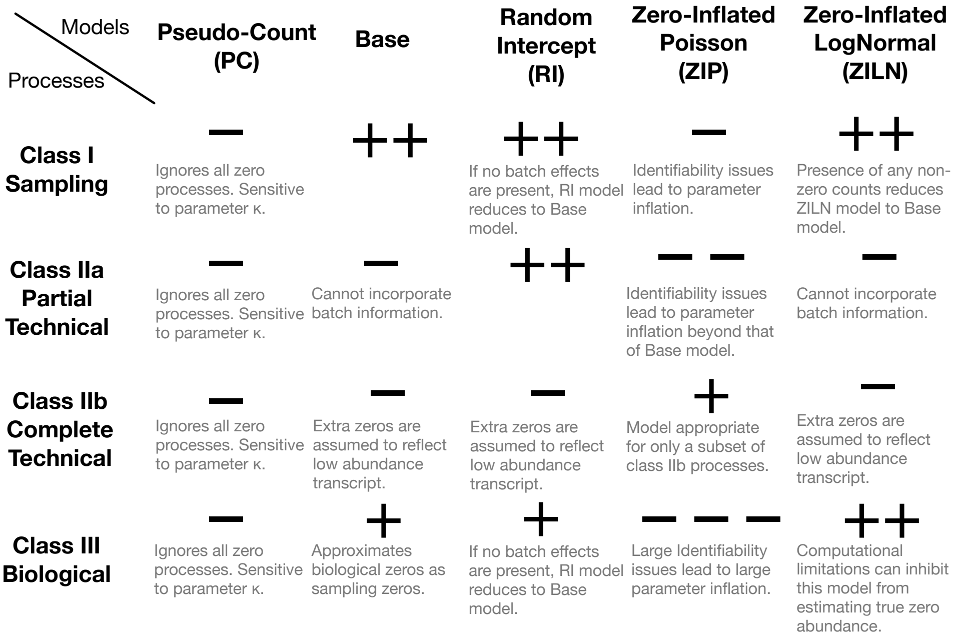
Gene	nZero (NPC,GW21)	$(\log_2 \frac{\text{NPC}}{\text{GW21}})_{\text{NB}}$	$(\log_2 \frac{\text{NPC}}{\text{GW21}})_{\text{ZINB}}$	$\text{DE}_{\text{NB}} - \text{DE}_{\text{ZINB}}$	ZIRM
SHISA2	(0/30,15/16)	5.2 (1)	0.86 (28)	4.34	4.00
TGFBR1	(2/30,7/16)	4.2 (4)	1.4 (19)	2.80	0.73
TRIM59	(0/30,11/16)	4.6 (3)	1.8 (11)	2.80	1.68
BIRC5	(0/30,11/16)	3.6 (10)	0.93 (26)	2.67	1.68
CENPF	(0/30,12/16)	3.9 (9)	1.5 (16)	2.40	2.00
SFRP1	(0/30,14/16)	4.1 (5)	1.7 (14)	2.40	3.00
KPNA2	(0/30,8/16)	4.6 (3)	2.3 (3)	2.30	1.00
NFIB	(7/30,0/16)	-5.4 (99)	-3.5 (95)	-1.90	-0.38
PDS5B	(0/30,9/16)	4 (6)	2.2 (4)	1.80	1.19
FZD3	(0/30,8/16)	3.5 (11)	1.7 (14)	1.80	1.00
CDV3	(3/30,6/16)	-0.68 (58)	0.12 (44)	-0.80	0.53
SC5D	(3/30,4/16)	1.1 (31)	-0.03 (51)	1.13	0.26
SERINC5	(2/30,5/16)	0.85 (33)	-0.07 (52)	0.92	0.44
IDH1	(2/30,3/16)	-0.8 (59)	0.03 (47)	-0.83	0.20
RELL1	(7/30,8/16)	0.58 (40)	-0.16 (55)	0.74	0.62
USP47	(2/30,8/16)	-0.17 (52)	0.63 (32)	-0.80	0.90
MIER1	(1/30,6/16)	-0.52 (56)	0.42 (38)	-0.94	0.63
TMEM33	(0/30,1/16)	-0.97 (63)	0.02 (49)	-0.98	0.09
SKIL	(2/30,4/16)	-1.2 (66)	0.22 (41)	-1.42	0.32
HACD3	(1/30,8/16)	0.66 (39)	-0.35 (58)	1.01	0.95

values to inform the estimated abundance of the same gene and thus the log-ratio of the mean count value for a gene including zero values would be a better approximation to the inferred differential expression from the NB model. The difference in differential expression correlated strongly with this ZIRM statistic over the 100 genes (Spearman ρ of 0.55, p-value of $3.4e^{-9}$, Figure S5). Thus this intuitive statistic, based on estimating DE with or without zero values, predicts the difference in DE between the NB and ZINB model in practice. This result further emphasizes that the concepts and intuition introduced in prior sections provides a useful framework for understanding results in real data scenarios.

4 Simulations

Motivated by the substantial and counter-intuitive impact of zero-inflation on modeling results in Section 3, we created a series of simulation studies designed to investigate the behavior of each model introduced in Section 2.2 on each zero generating process introduced in Section 2.1. We present only univariate simulations for ease of interpretation. The base, ZIP (Zero-Inflated Poisson), and ZILN (Zero-Inflated Log-Normal) models were applied to all 5 datasets to demonstrate the impact of mismatches between each zero generating processing and each model. In contrast, the PC (Pseudo-Count) and RI (Random Intercept) models were only applied where applicable or informative. For rhetorical purposes, hyper-parameter values were chosen to be applicable to each of the five simulations. The chosen hyper-parameters are as follows: $\sigma^2 = 3$, $\rho = -1$, $\tau^2 = 5$, $\nu = 0$, $\omega^2 = 2$, $\alpha = .5$, $\beta = .5$, $\zeta = 1$, and $\xi = 1$. A visual summary of the results of these 5 simulations is given in Figure 2.

All 5 of these simulations contain technical replicates in order to make the concepts discussed in Sections 2.1 and 2.2 clear. While the collection of technical replicates is uncommon in sequencing studies [30, 31], we felt inclusion of such replicates was the most straight forward way of identifying these models without relying on further modeling assumptions.



Models / Processes	Pseudo-Count (PC)	Base	Random Intercept (RI)	Zero-Inflated Poisson (ZIP)	Zero-Inflated LogNormal (ZILN)
Class I Sampling	− Ignores all zero processes. Sensitive to parameter κ.	++	++ If no batch effects are present, RI model reduces to Base model.	− Identifiability issues lead to parameter inflation.	++ Presence of any non-zero counts reduces ZILN model to Base model.
Class IIa Partial Technical	− Ignores all zero processes. Sensitive to parameter κ.	− Cannot incorporate batch information.	++	− − Identifiability issues lead to parameter inflation beyond that of Base model.	− Cannot incorporate batch information.
Class IIb Complete Technical	− Ignores all zero processes. Sensitive to parameter κ.	− Extra zeros are assumed to reflect low abundance transcript.	− Extra zeros are assumed to reflect low abundance transcript.	+ Model appropriate for only a subset of class IIb processes.	− Extra zeros are assumed to reflect low abundance transcript.
Class III Biological	− Ignores all zero processes. Sensitive to parameter κ.	+ Approximates biological zeros as sampling zeros.	+ If no batch effects are present, RI model reduces to Base model.	− − − Large Identifiability issues lead to large parameter inflation.	++ Computational limitations can inhibit this model from estimating true zero abundance.

Figure 2: A visual summary of the behavior of different models on different zero generating processes. This figure summarizes the simulations of Section 4. '+' endorse and '–' discourage the use of a given model in the presence of a given zero generating process.

4.1 Simulation 1: Highlighting Sampling Zeros

The first simulation consists of 5 random draws from a Poisson distribution with a rate parameter (λ) of 0.5. This simulation represents a single transcript within a single person measured with 5 technical replicates all processed in the same batch. The small value of λ ensured that the data would contain a number of class I (sampling) zeros with high probability. We applied the PC, base, ZIP, and ZILN models to this simulation. To demonstrate the impact of the choice of pseudo-count on the PC model, we applied the PC model with three different pseudo-counts (1, .5, and .05). We summarize and provide an intuitive explanation of the results (shown in Figure 3A) below:

PC model The PC model is quite sensitive to the choice of pseudo-count κ. Typical values for κ used in the analysis of sequence count data include .5, .65, and 1, as there is no generally optimal value that can be inferred from the observed data directly [33]. Here we found that only a choice of $\kappa = 0.05$ provided a close correspondence between the posterior mean of λ and the true simulated value of λ.

Base model The base model performs well, placing the posterior mean near the true simulated value of λ.

ZIP model Surprisingly, while the ZIP model is capable of modeling pure sampling zeros (*i.e.*, if $\theta_1 = 0$) this model demonstrates substantial inflation of λ compared to its true value. This occurs because the ZIP cannot distinguish as to whether zero values are due to a small value of λ (low abundance) and a small value of θ (low zero inflation) or a large value of λ (high abundance) and a large value of θ (high zero inflation). This interpretation is supported by a strong positive correlation in the posterior distribution of λ and θ shown in Figure S1. In addition, Figure S1 demonstrates that the regions of high posterior probability are spread out over a large range of possible λ and θ values. This uncertainty also appears in the long tails of the ZIP model's posterior distribution for λ. These results demonstrates the weak identifiability of zero-inflated models.

ZILN model The ZILN model performs nearly identical to the base model. To explain this behavior note that the presence of non-zero counts makes it extremely unlikely that the true value of λ is zero; if $\lambda = 0$ we would expect all counts to be zero. Therefore, the ZILN model

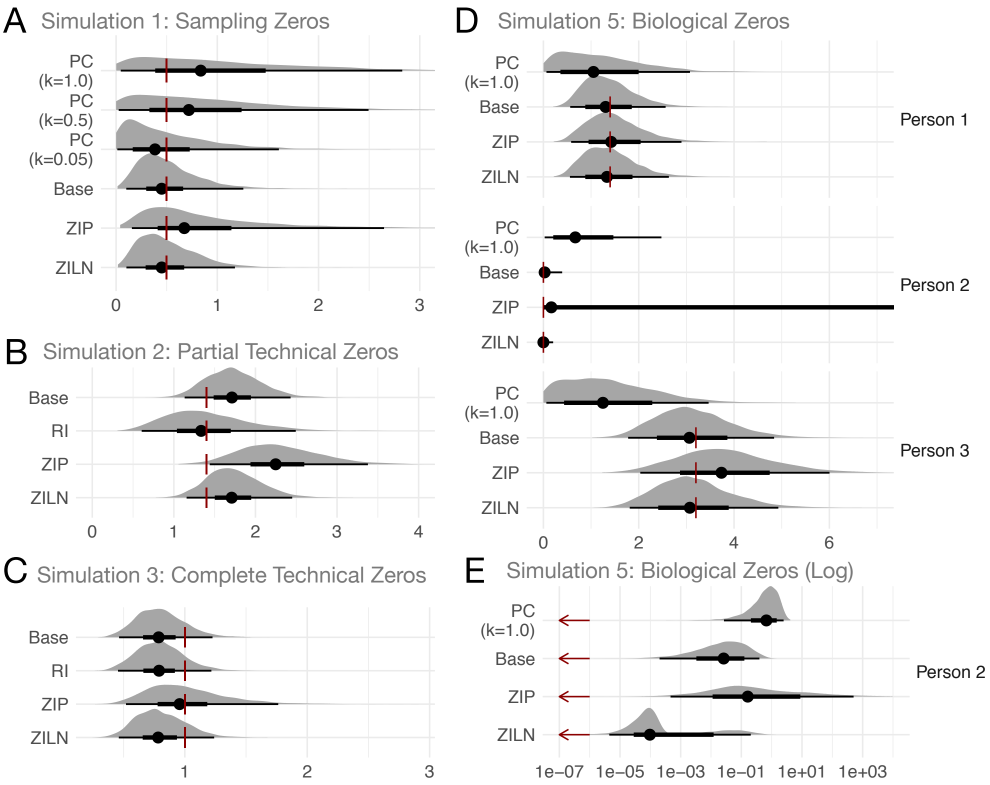


Figure 3: Posterior distribution of λ from each model applied to simulated datasets. Dark red vertical bar represents true value of λ. Posterior mean as well as the 66% and 95% credible intervals are shown in black. (A) Simulation 1 (sampling zeros), (B) Simulation 2 (partial technical zeros), (C) Simulation 3 (complete technical zeros) (D) Simulation 5 (biological zeros), (E) Person 2 from Simulation 5 shown on a log scale. PC, Pseudo-Count Model; RI, Random Intercept Model; ZIP, Zero-Inflated Poisson Model; ZILN, Zero-Inflated Log-Normal Model. Simulation 4 is a secondary example of a complete technical process and is shown in Figure S3

estimates that the true value of γ must be near zero. If $\gamma \approx 0$ then the ZILN model reduces to the base model thus explaining the similar behavior of these two models.

To investigate whether this notable behavior of the ZIP model could be due to small sample size (few technical replicates) in simulation 1, we repeated this analysis at a variety of sample size between 5 and 1280 each with the same rate parameters as above. For each sample size we simulated 30 datasets. For each simulated dataset both the base and ZIP models were fit. The distribution of the posterior means of each of these two models as a function of sample size is shown in Figure S2. We find that with increased sample size the inflation of λ does decrease, but even with 1280 samples per dataset, the ZIP model continues to demonstrate inflation of mean estimate of λ. In contrast, with only 5-10 samples, model I estimates λ near its true value without this bias.[5] This result demonstrates that the zero-inflated models can demonstrate bias even for extremely large sample sizes due to weak identifiability.

4.2 Simulation 2: Highlighting Partial Technical Zeros

The second simulation consists of 15 replicates samples split evenly into 3 batches with Poisson rate parameters 1.4, 0.6, and 3.2 respectively. This simulation represents a situation where polymerized chain reaction (PCR) efficiency varies by batch. We consider batch 1 to be derived from some gold standard measurement device that has no bias. As the rate parameters for each batch are all small, this dataset contains a mix of sampling and partial technical zeros. We summarize and provide an intuitive explanation of the results (shown in Figure 3B) below:

Base model The base model cannot incorporate batch information and therefore naively estimates that all 15 samples come from a distribution with a fixed rate parameter. In this way, the base model essentially estimates the true rate parameter as the mean of the rate parameters of the three batches. As this mean rate is higher than the true rate in batch 1, the base model inflates its abundance estimate compared to the true value[6].

RI model The RI model performs well in this simulation placing the posterior mean near the true value of λ.

ZIP model The posterior mean of the ZIP model lies even higher than that of the base or ZILN models. This may seem surprising given that the ZIP model can use batch information. This result can be understood in two parts. First, the ZIP cannot detect a shift in the overall Poisson rate parameter between batches, it can only detect differences in the rates of zeros between batches. This limitation causes the ZIP model to view the data, and inflate estimates, much as the base model does (based on the overall average rates between batches). Second, just as in prior simulations, the zero inflation component of the ZIP model essentially excludes some zero values from its estimates of λ and in doing so inflates the overall estimates for λ. Combining these two parts, the ZIP results can be seen as inflation through a similar mechanism as the base model plus even more inflation due to its the zero inflated component of the model.

ZILN model Here the ZILN model behaves identically to the base model. As in simulation 1, this occurs due to the presence of non-zero counts making it highly unlikely that the true $\lambda = 0$.

4.3 Simulations 3 and 4: Highlighting Complete Technical Zeros

The third simulation consists of 15 replicate samples from a Poisson distribution with rate parameter (λ) of 1. This simulation represents the following hypothetical situation: a single transcript is measured with technical replicates; however, each replicate has a 30% chance of catastrophic error causing a complete inability to measure that transcript. As with prior simulations, the small rate parameter ensures that the resulting simulated data contains sampling zeros in addition to

[5] That the ZIP model's bias improves with increasing sample size at all is due to the model using the variation of the non-zero counts to eventually approach the correct answer.

[6] Although the log-normal component of model I allow for over-dispersion compared to the Poisson distribution alone, the bias here is systematic and poorly modeled by such stochastic over-dispersion. As a result we would expect a negative binomial model to perform similarly to the RI model

complete technical zeros. We summarize and provide an intuitive explanation of the results (shown in Figure 3C) below:

Base model The base model underestimates the true simulated value of λ. This occurs because the base model incorrectly assumes the complete technical zeros are really sampling zeros. The excess zeros thus deflate the base models estimates of λ.

RI model Since all samples came from the same batch, there is no difference between the base and RI models. Thus we see that like the base model, the RI model underestimates teh true value of λ.

ZIP model The ZIP model performs well, placing the posterior mean of λ near its true simulated value.

ZILN model As in simulations 1 and 2, the presence of non-zero counts makes it highly unlikely that the true value of λ is near zero. Thus the non-zero counts force $\gamma \approx 0$ and the ZILN model reduces to the base model. This explains why the ZILN model performs identically to the base model.

The above simulation may seem contrived as it is unclear what experimental process would cause such a random but complete inability to measure a transcript within only select samples in a batch. For this reason, a second dataset of class IIb zeros was simulated (simulation 4). This simulation represents a single transcript measured in 15 replicate samples (5 replicates in each of 3 batches). However, due to the use of a different reagent or a missed experimental step, within batch 2 there is a complete lack of the transcript. 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. The results from simulation 4, which are shown in Figure S3, appear similar to those of simulation 3. The key difference is that here, the RI model performs better than the base or ZILN models but still underestimates the true value of λ. Surprisingly, here the ZIP model slightly over estimates the true simulated value of λ. These results of the RI and ZIP models again stem from each models inability to distinguish between which zeros are due to a sampling process and which are due to a technical process. Notably, these results demonstrate that the ZIP model performs well only in a subset of complete technical processes (*e.g.*, simulation 3) but many still cause over inflation of parameter estimates in other complete technical processes (*e.g.*, simulation 4).

4.4 Simulation 5: Highlighting Biological Zeros

The fifth simulation consists of 15 samples from three individuals (5 replicates each) with Poisson rate parameters 1.4, 0, and 3.2 respectively. This simulates a situation where the abundance of a single transcript is measured in three individuals of which two posses that transcript and one does not (biological zeros). As in the previous simulations, the small rate parameters ensure that this simulation contains sampling zeros as well as biological zeros. In addition, to simulate a situation in which an analyst naively chooses to model biological zeros with zero inflation, we consider a slight modification of equation (10) in the ZIP model where we replace θ_{x_i} with θ_{z_i}. This change reflects a change of modeling zero-inflation by batch to modeling zero-inflation by individual. We summarize and provide an intuitive explanation of the results (shown in Figure 3D and 3E) below:

PC model Unsurprisingly, the PC model performs poorly here, providing biased estimates in all three people[7].

Base model The base model performs quite well in this simulation. In the absence of any non-zero counts in person 2, the base model places posterior estimates of λ_2 on low values that would be expected to produce large numbers of sampling zeros.

ZIP model Most notably, the ZIP model massively overestimates value of which was so high that the posterior credible intervals were cropped in 3D to aid visualization of the other results. This behavior of the ZIP model can be understood through the same mechanism that caused the ZIP model to inflate parameter estimates in simulations 1, 2 and 4. Namely, the ZIP

[7] The PC model was included in this simulation to demonstrate how the inclusion of a fixed pseudo-count forces the posterior estimates for λ_2 to remain near the pseudo-count value without allowing the model to approach the true value of $\lambda_2 = 0$.

model has difficulty distinguishing between high λ_2 (high abundance) and high θ_2 (high zero inflation) versus low λ_2 (low abundance) and low θ_2 (low zero inflation). However, in contrast to previous simulations, here the difficulty is far more severe as all replicates from person 2 are zero and thus the ZIP model has no information to identify this model. This conclusion is supported by Figure S1 which demonstrates how the regions of highest posterior probability span both very high and very low values of θ_2 as the values of λ_2 vary over nearly 10 orders of magnitude. As we will show in Section 3, this feature of zero-inflated models can have profound impacts on modeling results in real data situations.

ZILN model The ZILN model performs well in this simulation estimate the true value of λ near its true value in all 3 people. To better understand the differences between the base and ZILN model results, the estimates for λ_2 are shown on a log scale in Figure 3E. Here the complication of biological zeros is emphasized as on a log scale, the true value of λ_2 is negative infinity. Notably, neither model is able to estimate this true value due to numerical precision limitations of computers in combination with our use of HMCMC which cannot handle a latent Dirac distribution but instead requires an approximating truncated normal distribution (*Methods*). Despite this, the zero inflation in the ZILN model does allow this model to estimate values of λ_2 approximately 2 orders of magnitude smaller than the base model. The ZILN model achieves this by placing significant posterior probability on large values of γ_2 which also gives this posterior estimate a distinctive bimodal shape. It is possible that had the ZILN model been inferred with an algorithm that allowed a latent Dirac distribution to be included (such as a Metropolis-within-Gibbs sampling scheme) the ZILN model would actually place non-negligible probability mass exactly on $\lambda_2 = 0$.

5 Discussion

Here we have argued that there are at least three types of zero values in sequence count data: under-detection zeros (class I), technical zeros (class II), and biological zeros (class III). We also argued that technical zeros could be subdivided based on whether they arise from partial (class IIa) or complete (class IIb) depletion of a transcript during sample processing. Additionally, we introduced a classification system of models intended to pair with our zero classification scheme. In Section 3 we demonstrated that modeling results on real data may differ significantly depending on the zero generating processes. Additionally, we demonstrated how the concepts of Section 2.1 and 2.2 enabled us to predict the ways in which the ZINB and NB models differed in terms of their estimates of differential expression. In Section 4 we further explored all pairwise relationships between the zero generating processes of Section 2.1 and the models of Section 2.2. Though these simulations we demonstrated how pseudo-count based methods (class 0 models) can introduce substantial biases into an analysis and can be sensitive to the choice of pseudo-count. We also demonstrated that mismatches between the choice of models and the zero generating process underlying the observed data can lead to spurious conclusions. In particular, we found that zero inflated models tend to inflate parameter estimates in the absence of complete technical processes and that this inflation remains even with sample sizes greater than 1000.

Of course, the degree to which different zero generating processes contribute to zero patterns in sequence count data is likely problem specific. For example, humans maintain a complex but consistent microbial community within their gut but with significant inter-individual variation in taxa profiles. Therefore a longitudinal analysis of a single individual microbiota may involve fewer class III zeros than a study investigating microbiota cross-sectionally over multiple individuals. Similarly, an analysis of human microbiota at the Phylum level will likely involve fewer class I zeros than a similar analysis conducted at the species level. As a final example, studies that use consistent experimental protocols with samples analyzed in a single batch will likely involve fewer class II zeros than a similar study where sample processing was conducted by different laboratories.

While the degree to which different zero generating processes are present in data is almost certainty problem specific, our results do suggest a number of potential pitfalls and recommendations for modeling sequence count data. Most notably, our results suggest that caution be exercised in the use of zero-inflated models. We demonstrated that these models can exhibit substantial bias due to weak identifiability and can lead to counter intuitive results in real data situations. While we cannot know whether these counter intuitive results are correct or not, we can instead ask whether the processes modeled by such models are plausible. In short, we believe it highly unlikely that complete technical processes are occurring to any substantial degree in real data. Most commonly zero-inflation models are used in single-cell RNA-seq studies to model so called "dropout" events [34, 35]. Such dropout events are described as resulting from a combination of stochastic sampling of low-abundance transcripts[34, 35]. This description of "dropouts" seems at odds with the complete technical process modeled by zero inflation and in fact seem to describe either a sampling or multivariate partial technical process (e.g., a competition to be counted). Taken together we recommend extreme caution be exercised in using zero inflated models in the analysis of sequence count data.

Beyond caution in the use of zero inflated models, our results suggest that biological zeros may be well modeled by zero adjusted model such as the base model which considers such zeros as arising due to a sampling process. Modeling biological absences as sampling zeros, relies on approximating true biological absences with very low (rather than absolute zero) abundance estimates. Two limitations of this approximation should be noted. First, such an approximation implicitly assumes that model structure (e.g., the dependencies between genes or taxa) depends smoothly on transcript abundances and does not display discontinuous behaviour based on presence/absence patterns. Yet such such an assumption may not be valid. For example, within a cell, the presence or absence of a given gene may disrupt regulatory pathways in a largely discontinuous manner. Second, modeling biological zeros as sampling zeros may not be computationally stable if performed in log space as a model flexible enough to allow inferred abundances to approach zero will equate to parameters tending towards negative infinity in log space. In contrast, class III models, such as the ZILN model, tend to be far more computationally intensive than unadjusted count model models (class I) and may, for different reasons be computationally unstable. Taken together we recommend that, in general, biological zeros be approximated as sampling zeros.

Overall we suggest that modeling a combination of sampling and partial technical zeros to be the most widely applicable approach to the analysis of sequence count data. With respect to partial technical zeros, we recommend models provide a means of accounting for batch variation and multivariate competition to be counted. The later, a competition to be counted, is less commonly addressed with modeling but may be accounted for though the use of multinomial based models. In particular, we believe this approach, which makes use of a sampling approximation to biological zeros, addresses the most common sources of zeros likely present in real data. Notable available methods that meet these criteria include the multinomial logistic normal dynamic linear models of Silverman et al. [30] and the multinomial logistic normal linear models of Grantham et al. [20] and [19].

While we believe the concepts and classification systems that we have introduced provide a useful framework for directing modeling choices and model inspection, two caveats should be mentioned. First, we do not believe that our classification systems encompass the entire complexity of zero generating processes nor modeling approaches. For example, in single cell RNA-seq not only may there be a competition to be counted between transcripts, but zeros may arise due to a competition to be counted between entire cells. While such a competition to be counted between cells can be thought of in terms of a partial technical process, further discussion is needed to fully appreciate the complexity of this type of zero generating process. Second, our prescriptive recommendations should only be taken as loose recommendations as specific features of a given study should ultimately drive modeling decisions. For example, while we generally recommend against the use of zero-inflation in modeling sequence count data, such models may be called for in select situations. Consider a single cell RNA-seq study focuses on a set of genes that are only expressed at certain phase of the cell cycle. If the aim is to investigate the expression of this gene when it is expressed, ignoring cases when it is not expressed, then the use of a zero-inflated model may be appropriate even though this represents a zero value due to a biological process rather than a complete technical process. Despite these caveats we believe that both our classification of zero values and our classification of models will provide a useful framework for model design and inspection.

Ultimately, the difficulty posed by zero values stems from the fact that zeros represent a lack of information that require experimental advances to fully resolve. While a count of 100 represents strong support for the presence of a transcript with an abundance between 99 and 101, a zero count represents only that that transcript was not observed without informing as to the cause of non-observation. Assumptions regarding the possible processes leading to non-observation can aid in interpreting zero values. For example, assuming the presence of only sampling or biological zero generating processes, a zero value may be interpreted as a value smaller than 1. In contrast, assuming the presence of both sampling and technical zero generating processes, a zero value could represent a value greater or smaller than 1. Unfortunately, it is unlikely that the validity of such assumptions can be assessed through the use of sequence counting alone; instead, such evaluation will likely require external experimental validation. For example, in microbiome studies resolving whether a zero value stems from a biological or a sampling / technical process may require selective growth experiments or more targeted molecular assays designed specifically to evaluate presence absence of a select taxa. In this way we believe that the problem of zero values represents a fundamental experimental limitation that requires further experimental advances to resolve. In the meantime, we believe that the handling of zero values should be driven by well described assumptions regarding the possible processes generating zero values.

6 Methods

6.1 Data Simulation

For each simulation, data was generated as described in Section 4. Notably, to aid in interpretation of model outputs, simulations that had low likelihood under the simulating model were repeated. This procedures was performed to ensure that each simulated dataset contained the necessary information for the true parameter values to be recovered. This was done for simulations in Figure 3 not for simulations in Figure S2.

6.2 Posterior Inference

For readability, all 5 models were implemented in the Stan modeling language which makes use of Hamiltonian Monte Carlo (HMC) sampling [36]. Model inference was performed using 4 parallel chains each with 1000 transitions for warmup and adaptation and 1000 iterations collected as posterior samples. Convergence of chains was determined by by manual inspection of sampler trace plots and through inspection of the split $\hat{R}$ statistic.

Inference of the ZILN model was modified from the form given in Section 2.2.5 due to the difficulty of representing the latent Dirac distribution using the Stan modeling language. Instead, the Dirac distribution in the ZILN model was approximated with a truncated normal distribution with mean 0 and variance 0.0001.

6.3 Analysis of Single Cell RNA-Seq Data

Data analysis was performed as described in Section 3. The ZINB model was differentiated from the NB model using the parameter *epsilon_min_logit*. Briefly, larger values of this parameter more strongly penalizes zero inflation in the model. The ZINB model used the default value for *epsilon_min_logit* to allow zero inflation, whereas the NB model used a value of 10^{14} to ensure no zero inflation would be used.

6.4 Code availability

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

References

[1] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, 2010.

[2] J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer, and R. Knight, "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, 2011.

[3] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, p. 550, 2014.

[4] A. T. L. Lun, K. Bach, and J. C. Marioni, "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts," *Genome Biology*, vol. 17, no. 1, p. 75, 2016.

[5] M. Dziubianau, J. Hecht, L. Kuchenbecker, A. Sattler, U. Stervbo, C. Rödelsperger, P. Nickel, A. U. Neumann, P. N. Robinson, S. Mundlos, H.-D. Volk, A. Thiel, P. Reinke, and N. Babel, "TCR Repertoire Analysis by Next Generation Sequencing Allows Complex Differential Diagnosis of T Cell-Related Pathology," *American Journal of Transplantation*, vol. 13, no. 11, pp. 2842–2854, 2013.

[6] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing," *Nature Methods*, vol. 4, no. 8, pp. 651–657, 2007.

[7] D. D. Licatalosi, A. Mele, J. J. Fak, J. Ule, M. Kayikci, S. W. Chi, T. A. Clark, A. C. Schweitzer, J. E. Blume, X. Wang, J. C. Darnell, and R. B. Darnell, "HITS-CLIP yields genome-wide insights into brain alternative RNA processing," *Nature*, vol. 456, no. 7221, pp. 464–469, 2008.

[8] S. Weiss, Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham, E. R. Hyde, and R. Knight, "Normalization and microbial differential abundance strategies depend upon data characteristics," *Microbiome*, vol. 5, no. 1, p. 27, 2017.

[9] A. Kaul, S. Mandal, O. Davidov, and S. D. Peddada, "Analysis of Microbiome Data in the Presence of Excess Zeros," *Frontiers in Microbiology*, vol. 8, p. 2114, 2017.

[10] J. D. Silverman, A. D. Washburne, S. Mukherjee, and L. A. David, "A phylogenetic transform enhances analysis of compositional microbiota data," *eLife*, vol. 6, p. e21887, 2017.

[11] H. Li, "Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis," *Annual Review of Statistics and Its Application*, vol. 2, no. 1, pp. 73–94, 2015.

[12] G. B. Gloor, J. M. Macklaim, M. Vu, and A. D. Fernandes, "Compositional uncertainty should not be ignored in high-throughput sequencing data analysis," *Austrian Journal of Statistics*, vol. 45, no. 4, p. 73, 2016.

[13] L. Xu, A. D. Paterson, W. Turpin, and W. Xu, "Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data," *PLOS ONE*, vol. 10, no. 7, p. e0129606, 2015.

[14] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado, *Modeling and analysis of compositional data*. Statistics in practice, John Wiley & Sons, Ltd, 2015.

[15] S. Sun, M. Hood, L. Scott, Q. Peng, S. Mukherjee, J. Tung, and X. Zhou, "Differential expression analysis for RNAseq using Poisson mixed models.," *Nucleic acids research*, vol. 45, no. 11, p. e106, 2017.

[16] P. J. McMurdie and S. Holmes, "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible," *PLoS Computational Biology*, vol. 10, no. 4, 2014.

[17] P. S. La Rosa, J. P. Brooks, E. Deych, E. L. Boone, D. J. Edwards, Q. Wang, E. Sodergren, G. Weinstock, and W. D. Shannon, "Hypothesis testing and power calculations for taxonomic-based human microbiome data," *PLOS ONE*, vol. 7, no. 12, pp. 1–13, 2012.

[18] T. Ijö, C. L. Mü Ller, and R. Bonneau, "Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing," *Bioinformatics*, 2017.

[19] F. Xia, J. Chen, W. K. Fung, and H. Li, "A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis," *Biometrics*, vol. 69, no. 4, pp. 1053–1063, 2013.

[20] N. S. Grantham, B. J. Reich, E. T. Borer, and K. Gross, "MIMIX: a Bayesian Mixed-Effects Model for Microbiome Data from Designed Experiments," *arXiv*, 2017.

[21] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert, "A general and flexible method for signal extraction from single-cell RNA-seq data," *Nature Communications*, vol. 9, no. 1, p. 284, 2018.

[22] Z. Li, K. Lee, M. R. Karagas, J. C. Madan, A. G. Hoen, A. James O 'malley, and H. Li, "Conditional regression based on a multivariate zero-inflated logistic normal model for microbiome relative abundance data," *arXiv*, 2017.

[23] K. H. Lee, B. A. Coull, A.-B. Moscicki, B. J. Paster, and J. R. Starr, "Bayesian Variable Selection for Multivariate Zero-Inflated Models: Application to Microbiome Count Data," *arXiv*, 2017.

[24] G. Finak, A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, and R. Gottardo, "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data.," *Genome biology*, vol. 16, p. 278, 2015.

[25] A. Kaul, O. Davidov, and S. D. Peddada, "Structural zeros in high-dimensional data with applications to microbiome studies," *Biostatistics*, vol. 18, no. 3, 2017.

[26] M. Farris and J. Olson, "Detection of actinobacteria cultivated from environmental samples reveals bias in universal primers," *Letters in Applied Microbiology*, vol. 45, no. 4, pp. 376–381, 2007.

[27] P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad, "Batch effects and the effective design of single-cell gene expression studies.," *Scientific reports*, vol. 7, p. 39921, 2017.

[28] N. Li, D. A. Elashoff, W. A. Robbins, and L. Xun, "A hierarchical zero-inflated log-normal model for skewed responses," *Statistical Methods in Medical Research*, vol. 20, no. 3, pp. 175–189, 2011.

[29] C. M. Carvalho, N. G. Polson, and J. G. Scott, "Handling sparsity via the horseshoe," in *Artificial Intelligence and Statistics*, pp. 73–80, 2009.

[30] J. D. Silverman, H. K. Durand, R. J. Bloom, S. Mukherjee, and L. A. David, "Dynamic linear models guide design and analysis of microbiota studies within artificial human guts," *Microbiome*, vol. 6, no. 1, p. 202, 2018.

[31] J. D. Silverman, L. Shenhav, E. A. Halperin, S. A. Mukherjee, and L. A. David, "Statistical considerations in the design and analysis of longitudinal microbiome studies," *bioRxiv*, 2018.

[32] A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, N. Ramalingam, G. Sun, M. Thu, M. Norris, R. Lebofsky, D. Toppani, D. W. Kemp, M. Wong, B. Clerkson, B. N. Jones, S. Wu, L. Knutsson, B. Alvarado, J. Wang, L. S. Weaver, A. P. May, R. C. Jones, M. A. Unger, A. R. Kriegstein, and J. A. A. West, "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex," *Nature Biotechnology*, vol. 32, no. 10, pp. 1053–1058, 2014.

[33] J. Aitchison, *The statistical analysis of compositional data.* Monographs on statistics and applied probability, London ; New York: Chapman and Hall, 1986.

[34] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature methods*, vol. 11, no. 7, p. 740, 2014.

[35] W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, and D. J. Garry, "Drimpute: imputing dropout events in single cell rna sequencing data," *BMC Bioinformatics*, vol. 19, no. 1, p. 220, 2018.

[36] A. Gelman, D. Lee, and J. Guo, "Stan: A probabilistic programming language for bayesian inference and optimization," *Journal of Educational and Behavioral Statistics*, vol. 40, no. 5, pp. 530–543, 2015.

Supplementary Material

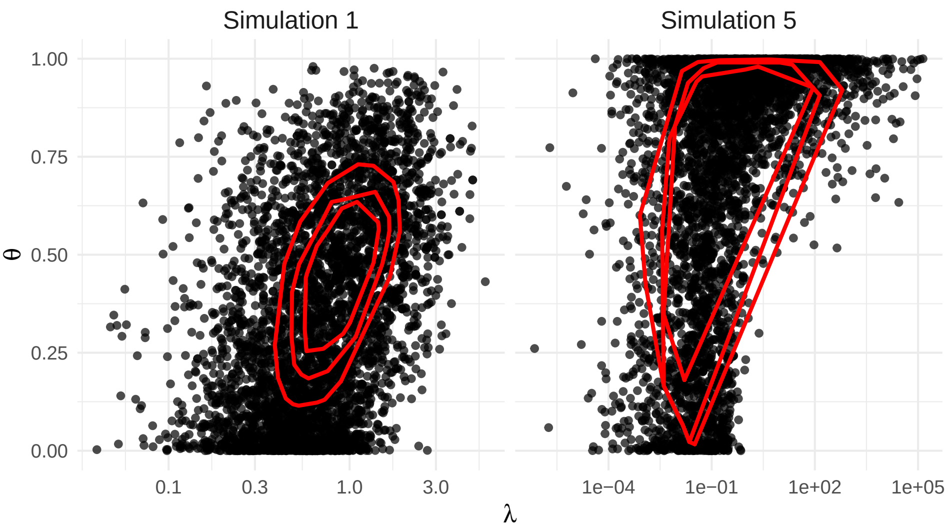
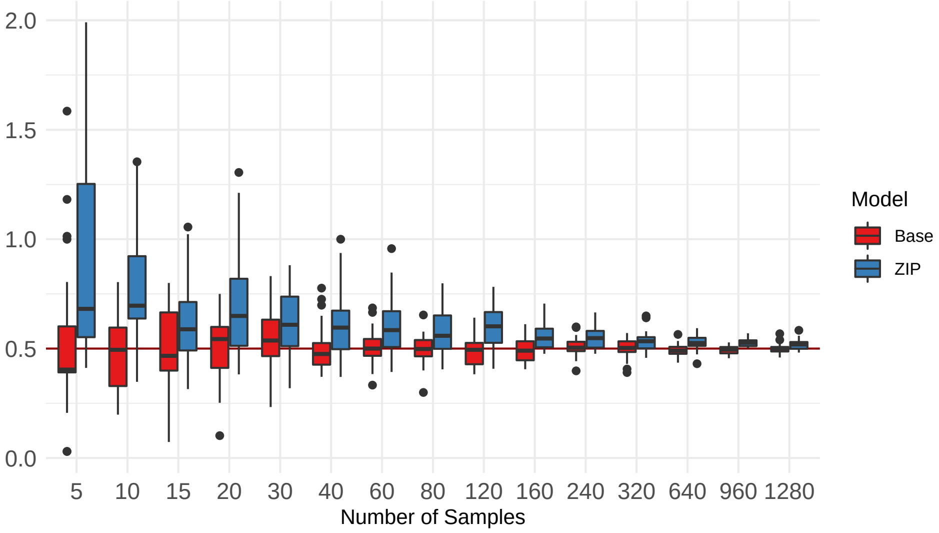


Figure S1: Posterior samples of λ and θ for the ZIP model applied to simulation 1 (sampling zeros) and simulation 5 (biological zeros). For simulation 5, the posterior distribution is of λ_2 and θ_2. The 80%, 90%, and 95% highest posterior density regions for the log posterior probability are shown in red.



]

Figure S2: Distribution of posterior mean estimates of λ from the Base and ZIP models for 30 datasets simulated as in Simulation 1 (sampling zeros; Poisson with rate parameter of 1) but with varying numbers of samples.

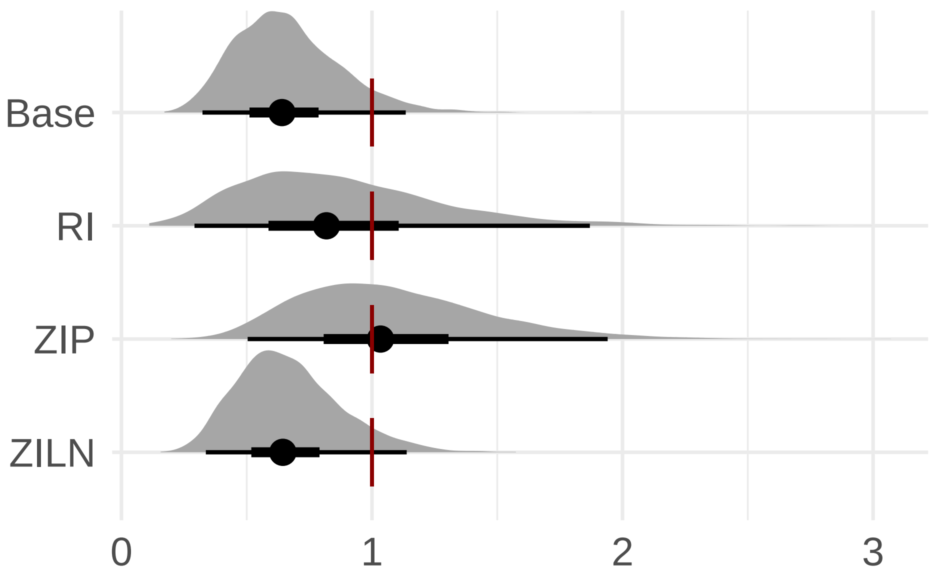


Figure S3: Posterior distribution of λ from each model applied to simulation 4 (second example of complete technical zeros). Dark red vertical bar represents true value of λ. Posterior mean as well as the 66% and 95% credible intervals are shown in black.

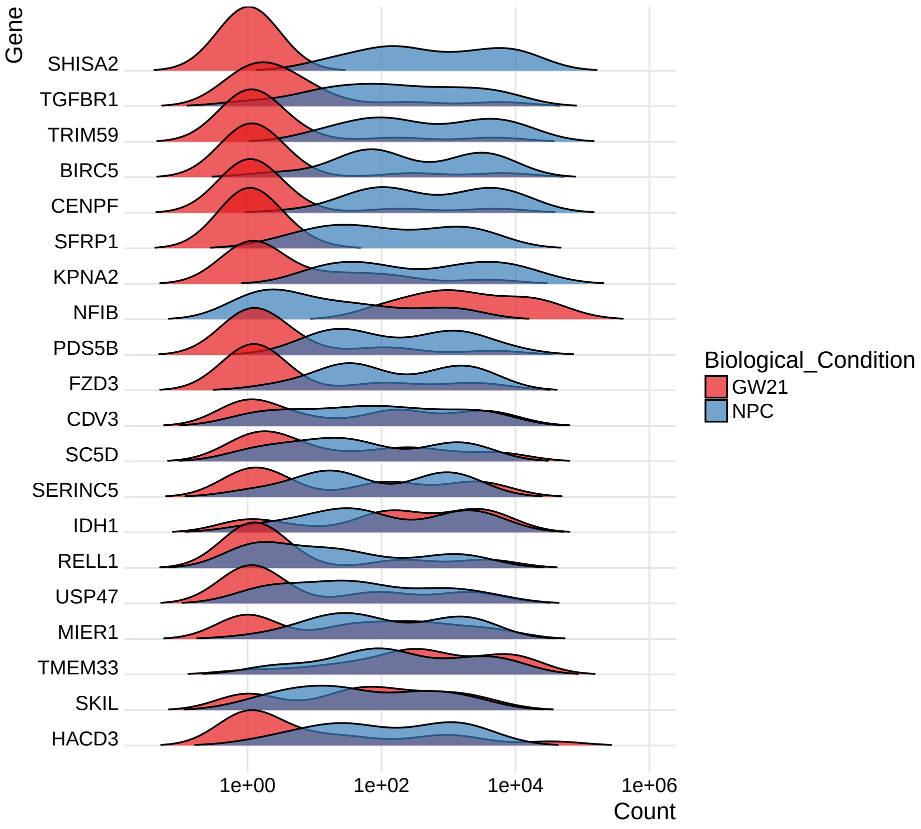


Figure S4: Kernel Density estimates for the distribution of counts of the 20 highlighted genes in Table 1.

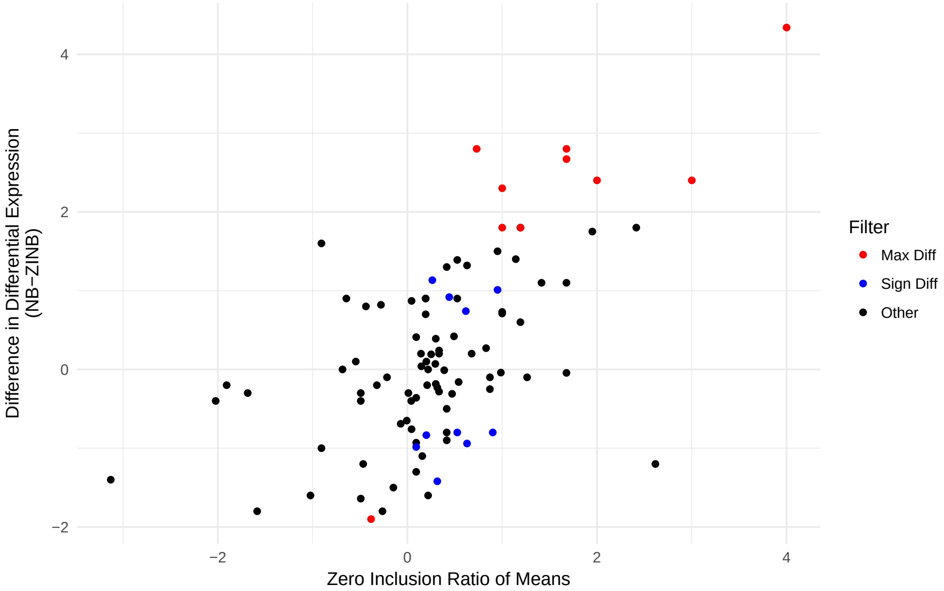


Figure S5: The observed difference in differential expression estimates for each of the 100 genes analyzed in section 3 correlates with the zero inclusion ratio of means statistic. The 10 genes with the largest absolute difference in estimated differential expression between the ZINB and NB models are shown in red and correspond to the first 10 genes in table 1. The 10 genes found to have differing signs of estimated differential expression between the ZINB and NB models are shown in blue and correspond to the last 10 genes in table 1.