Supplement: Naught all zeros in sequence count data are the same

Justin D. Silverman[1,2,3], Kimberly Roche[4], Sayan Mukherjee[4,5,6], and Lawrence A. David[4,6,7]

[1] *College of Information Science and Technology, Pennsylvania State University, State College, PA, 16802*
[2] *Institute for Computational and Data Science, Pennsylvania State University, State College, PA, 16802*
[3] *Department of Medicine, Pennsylvania State University, Hershey, PA, 17033*
[4] *Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27708*
[5] *Departments of Statistical Science, Mathematics, Computer Science, Biostatistics & Bioinformatics, Duke University, Durham, NC 27708*
[6] *Center for Genomic and Computational Biology, Duke University, Durham, NC 27708*
[7] *Department of Molecular Genetics and Microbiology, Duke University, Durham, NC 27708*
[6] *Denotes Co-corresponding Authors*

Contents

List of Figures

1 Detailed Description of Simulation Results

Here we extend the discussion in the main text and give a detailed description of our simulation results. We want to intuitively explain why these results appear the way they do.

1.1 Simulation 1: Highlighting Sampling Zeros

The first simulation consists of five 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 sampling zeros with high probability. We applied the PC, Base, ZIP, and BZ 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 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 we cannot directly infer a generally optimal value from the observed data [1]. Here we found that $\kappa = 0.05$ 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 While the ZIP model is capable of modeling pure sampling zeros (*i.e.*, if $\theta_1 = 0$), this model substantial inflated λ compared to its true value. The ZIP cannot distinguish between zero values due to low abundance and low zero inflation (small λ and small θ) and zero values due to high abundance and high zero inflation (large λ and large θ). This interpretation is supported by a strong positive correlation in the posterior distribution of λ and θ shown in Figure S10. Figure S10 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.

BZ model The BZ model performs nearly identically to the base model. 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. The BZ model estimates that the true value of γ must be near zero. If $\gamma \approx 0$ then the BZ model reduces to the base model.

We repeated this analysis at a variety of sample size between 5 and 1280 with the same rate parameters as above. For each sample size we simulated 30 datasets. For each simulated dataset, we fit both the base and ZIP models. The distribution of the posterior means of each of these two models as a function of sample size is shown in Figure S11. With increased sample size, the inflation of λ decreases, 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, the base Model estimates λ near its true value.[1] Thus estimates from zero-inflated models can demonstrate bias even for extremely large sample sizes.

1.2 Simulation 2: Highlighting Batch-Specifc 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. 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):

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. The base model estimates the rate parameter as the mean of the rate parameters of the three batches. As this mean rate is higher than the batch 1 rate, the base model inflates its abundance estimate.

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

[1] That the ZIP model's biased estimates improve with increasing sample size at all is because the model uses the variation of the non-zero counts to eventually approach the correct answer.

ZIP model The posterior mean of the ZIP model lies higher than that of the base or BZ models. This may seem surprising because 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, like the base model does, based on the overall average rates between batches. Second, the zero inflation component of the ZIP model 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 as the base model does, plus more inflation due to its zero inflated component.

BZ model Here the BZ 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 $\lambda = 0$.

1.3 Simulations 3 and 4: Highlighting Sample-Specific and Batch-Specific Complete Technical Zeros

The third simulation consists of 15 replicate samples from a Poisson distribution with rate parameter λ of 1. This simulation represents a hypothetical situation: a single transcript is measured with technical replicates; 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 data contains sampling zeros and complete technical zeros. We summarize and provide an intuitive explanation of the results (shown in Figure 3C):

Base model The base model underestimates λ. The base model incorrectly assumes the complete technical zeros are really sampling zeros. The excess zeros thus deflate the base model's estimates of λ.

RI model Since all samples came from the same batch, there is no difference between the base and RI models. So the RI model also underestimates the true value of λ.

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

BZ 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. The non-zero counts force $\gamma \approx 0$ and the BZ model reduces to the base model. This explains why the BZ model performs identically to the base model.

This simulation may be unrealistic, as it is unclear what experiment would cause a random but complete inability to measure a transcript within only select samples in a batch (sample-specific). So we simulated a second dataset of batch-specific complete technical zeros. In simulation 4, a single transcript is 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 appear similar to those of simulation 3. The difference is that the RI model performs better than the base or BZ models but still underestimates the true value of λ. The ZIP model slightly overestimates λ. These results of the RI and ZIP models stem from each model's inability to distinguish between which zeros are due to a sampling process and which are due to a technical process. The ZIP model performs well only in a subset of complete technical processes, *e.g.*, simulation 3, but may still cause over-inflation of parameter estimates in other complete technical processes (*e.g.*, simulation 4).

1.4 Simulation 5: Highlighting Biological Zeros

The fifth simulation consists of 15 samples from three individuals with Poisson rate parameters 1.4, 0, and 3.2. This simulates a situation where the abundance of a single transcript is measured in three individuals: two possess that transcript and one does not. As in the previous simulations, the small rate parameters ensure that this simulation contains sampling zeros as well as biological

zeros. To simulate a situation in which the ZIP model is used in a condition-specific way, we modify the ZIP model by replacing θ_{x_i} with θ_{z_i}. This changes modeling zero-inflation by batch to modeling zero-inflation by individual. We summarize and provide an intuitive explanation of the results (shown in Figure 3E and S9):

PC model The PC model performs poorly, providing biased estimates in all three people[2].

Base model The base model performs well in this simulation. With no 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 The ZIP model massively overestimates value of λ_2 which was so high that the posterior credible intervals were cropped in 2E to aid visualization of the other results. This behavior of the ZIP model comes from the same mechanism that inflated parameter estimates in simulations 1, 2 and 4. Namely, the ZIP model has difficulty distinguishing between high abundance and high zero inflation (high λ_2 and high θ_2) and low abundance and low zero inflation (low θ_2 and low λ_2). 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 S10 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.

BZ model The BZ model performs well in this simulation and estimates λ well in all 3 people. To see the differences between the base and BZ model results, the estimates for λ_2 are shown on a log scale in Figure S9. The complication of biological zeros is emphasized as on a log scale, the true value of λ_2 is negative infinity. Neither model can estimate this true value due to numerical precision limitations of computers and our use of HMCMC, which cannot handle a latent Dirac distribution and requires an approximating truncated normal distribution (*Methods*). But the zero inflation in the BZ model estimates values of λ_2 approximately two orders of magnitude smaller than the base model. The BZ model places significant posterior probability on large values of γ_2 which also gives this posterior estimate a distinctive bimodal shape. If we had inferred the BZ model with an algorithm that included a latent Dirac distribution, such as a Metropolis-within-Gibbs sampling scheme, the BZ model might place non-negligible probability mass exactly on $\lambda_2 = 0$.

2 Non-Condition-Specific Zero-Inflation

To investigate whether our results using the ZINB model were unique to condition-specific zero-inflation models, we repeated our analysis with a ZINB model that was fixed to only infer non-condition-specific zero-inflation (see Section 6 for more details). While we find less discrepancy between the NB and non-condition-specific ZINB model (ncsZINB), the observed patterns are similar with an average discrepancy of 32.7% (range: 3.0%-48.0%) among the top-50 most differentially expressed sequences. Similarly, for the top-5 most differentially expressed sequences the average disagreement averaged 23.0% and reached 60.0% for one dataset (Figures S3-S5). In parallel to the condition-specific case, we again observe a strong correlation between the difference in inferred differential expression between the ncsZINB and NB models and the inferred zero-inflation (absolute value of Spearman rho > 0.08 and p-value ≈ 0 for all 6 datasets; Figure S3). That is, even in the absence of condition-specific zero-inflation, the ncsZINB model interpreted that zeros in presence-absence-like cases were evidence of high levels of zero inflation rather than evidence of differential expression.

[2]We included the PC model to show how including 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 $\lambda_2 = 0$.

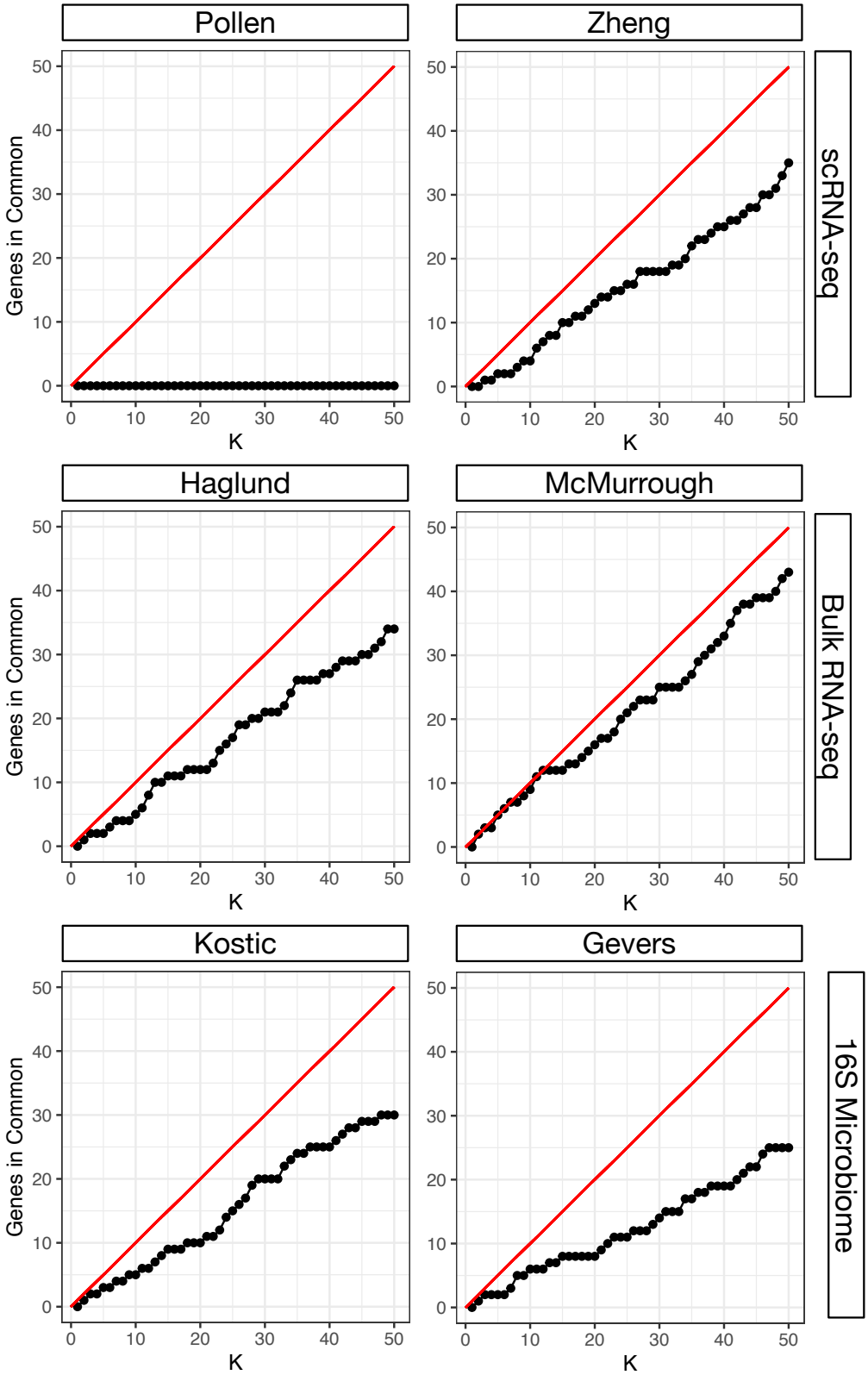

Figure S1: The ZINB and NB models often disagree regarding which sequences are the most differential expressed. For each dataset the intersection between the top-K most deferentially expressed sequences according to the NB and ZINB models is shown as a function of K.

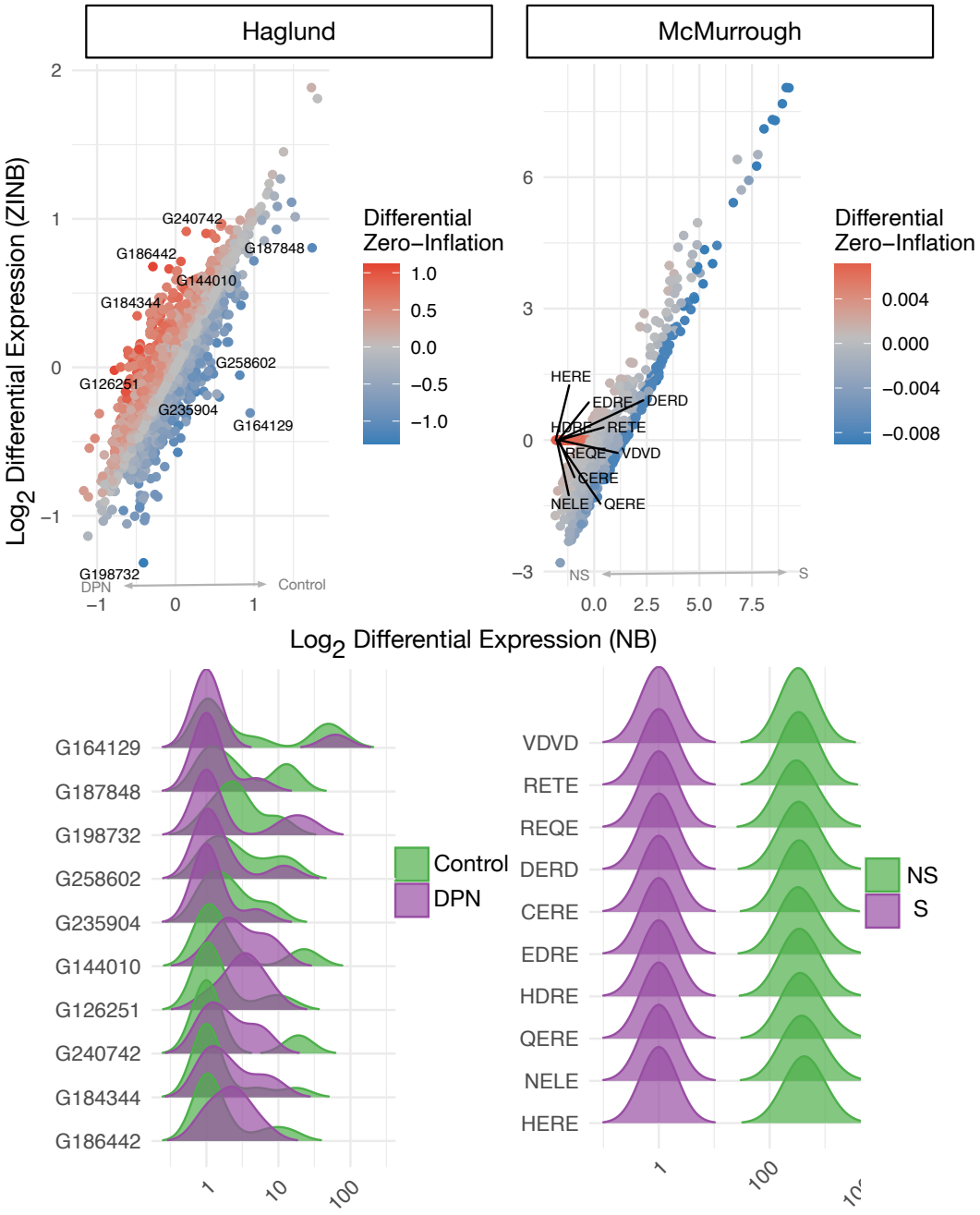

Figure S2: Log base 2 differential expression for the ZINB and NB models are shown after each was applied to two different bulk RNA-seq datasets. Dots represent different genes, and each is colored according to the degree of differential zero-inflation as estimated by the ZINB model. For each dataset, 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.

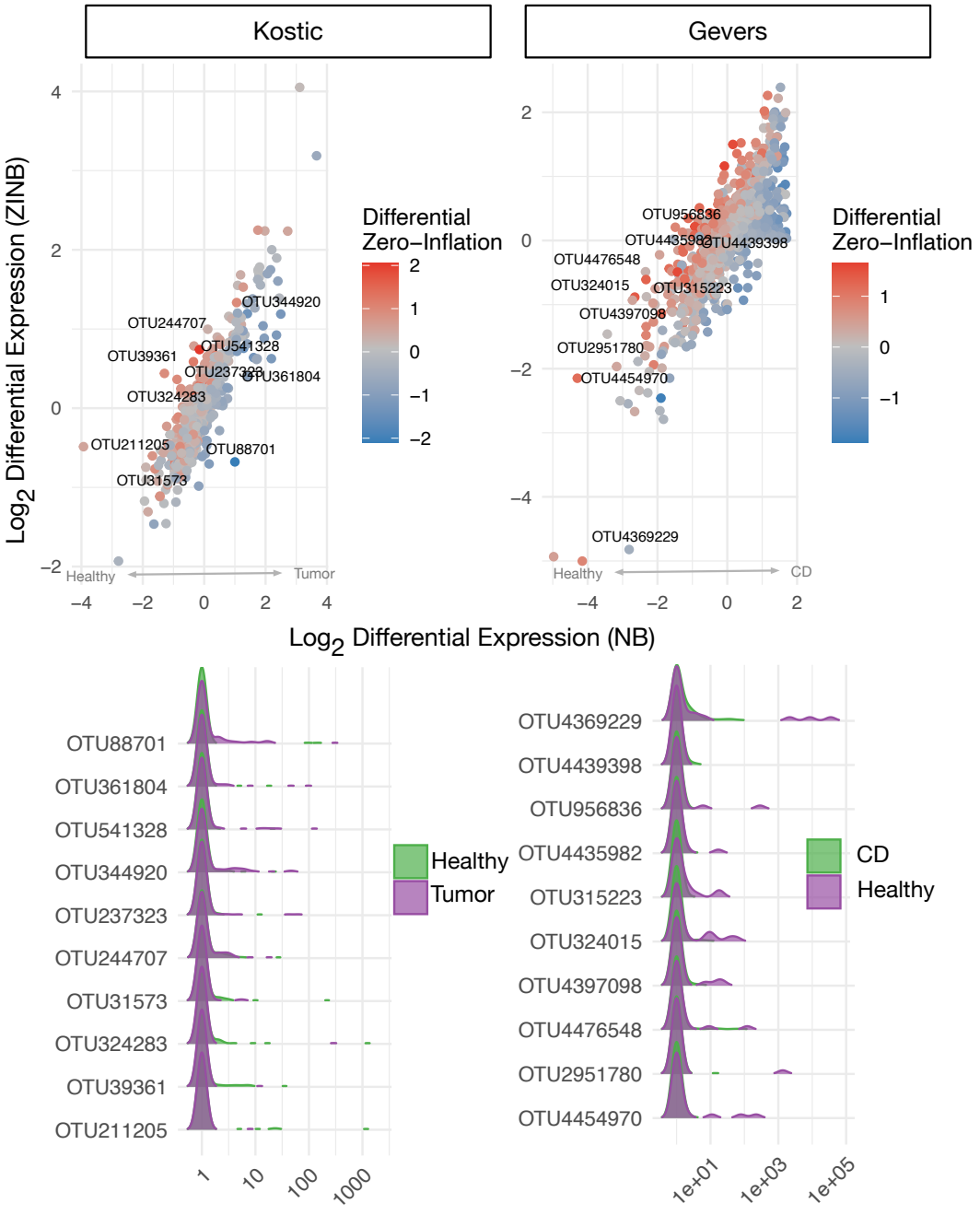

Figure S3: Log base 2 differential expression for the ZINB and NB models are shown after each was applied to two different 16S rRNA surveys. Dots represent different taxa, and each is colored according to the degree of differential zero-inflation as estimated by the ZINB model. For each dataset, the 10 taxa that have the largest discrepancy between inferred DE are labeled and their distribution is in each condition is plotted in the bottom panel.

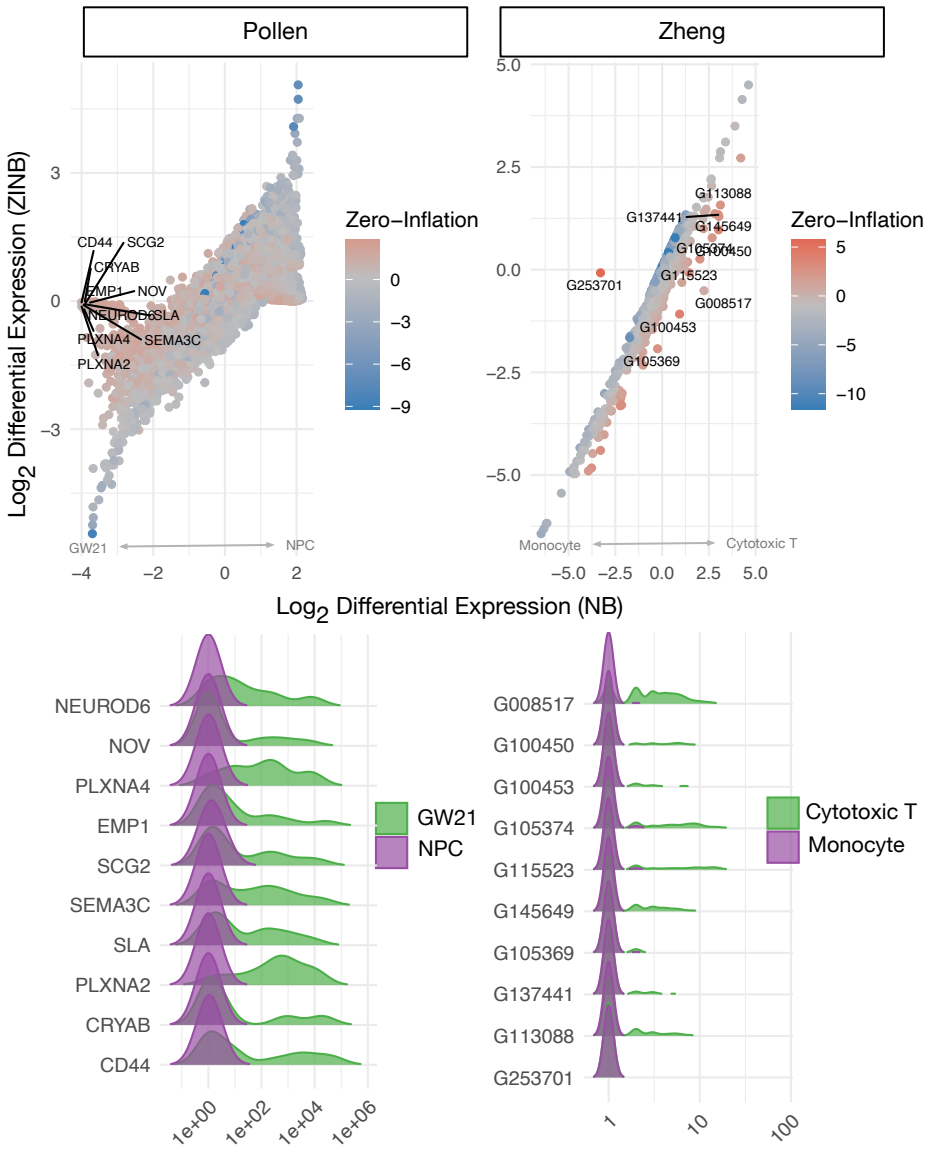

Figure S4: Differential expression (DE) estimates from a negative binomial (NB) and non-condition-specific zero-inflated negative binomial (ncsZINB) model can differ substantially. Log base 2 differential expression for the ncsZINB and NB models are shown after each was applied to two different single cell RNA-seq datasets. Dots represent different genes, and each is colored according to the degree of zero-inflation as estimated by the ncsZINB 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.

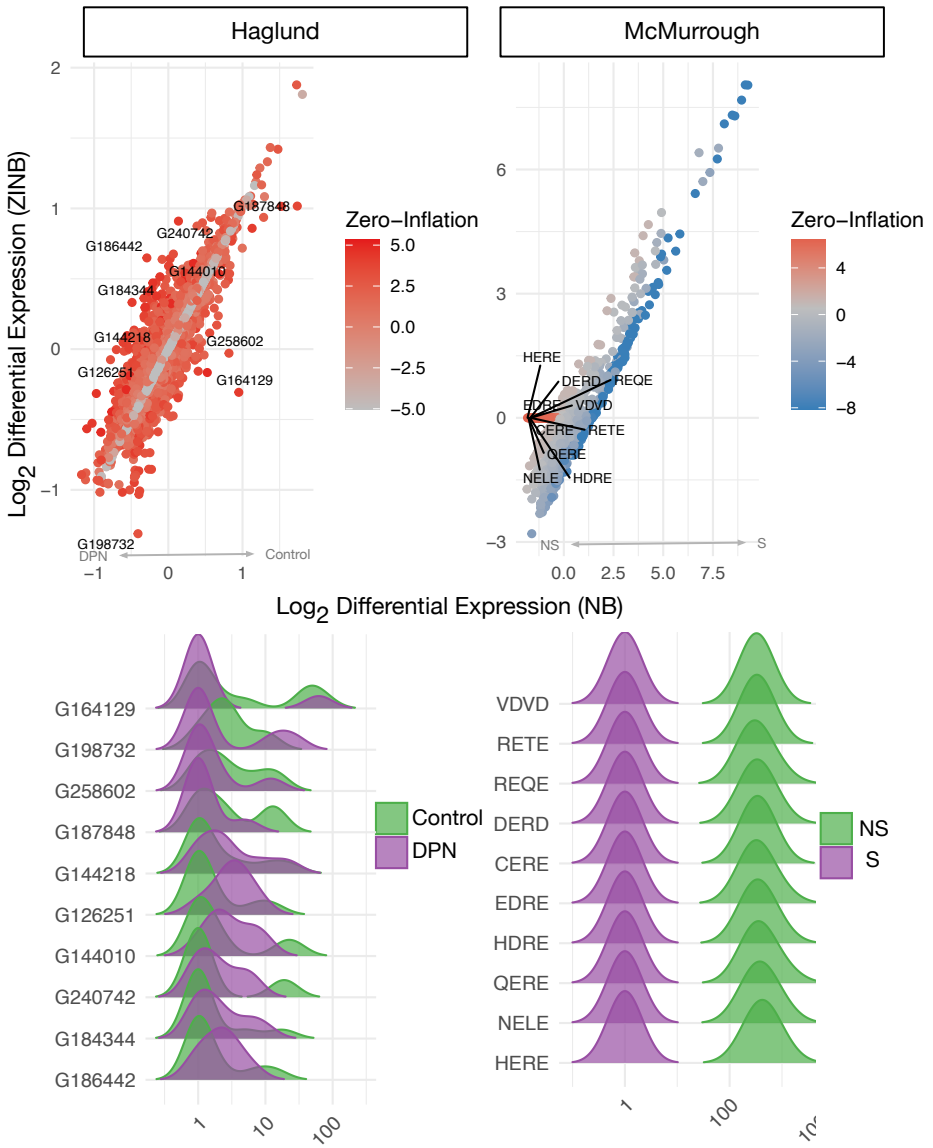

Figure S5: Differential expression (DE) estimates from a negative binomial (NB) and non-condition-specific zero-inflated negative binomial (ncsZINB) model can differ substantially. Log base 2 differential expression for the ncsZINB and NB models are shown after each was applied to two different bulk RNA-seq datasets. Dots represent different genes, and each is colored according to the degree of zero-inflation as estimated by the ncsZINB 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.

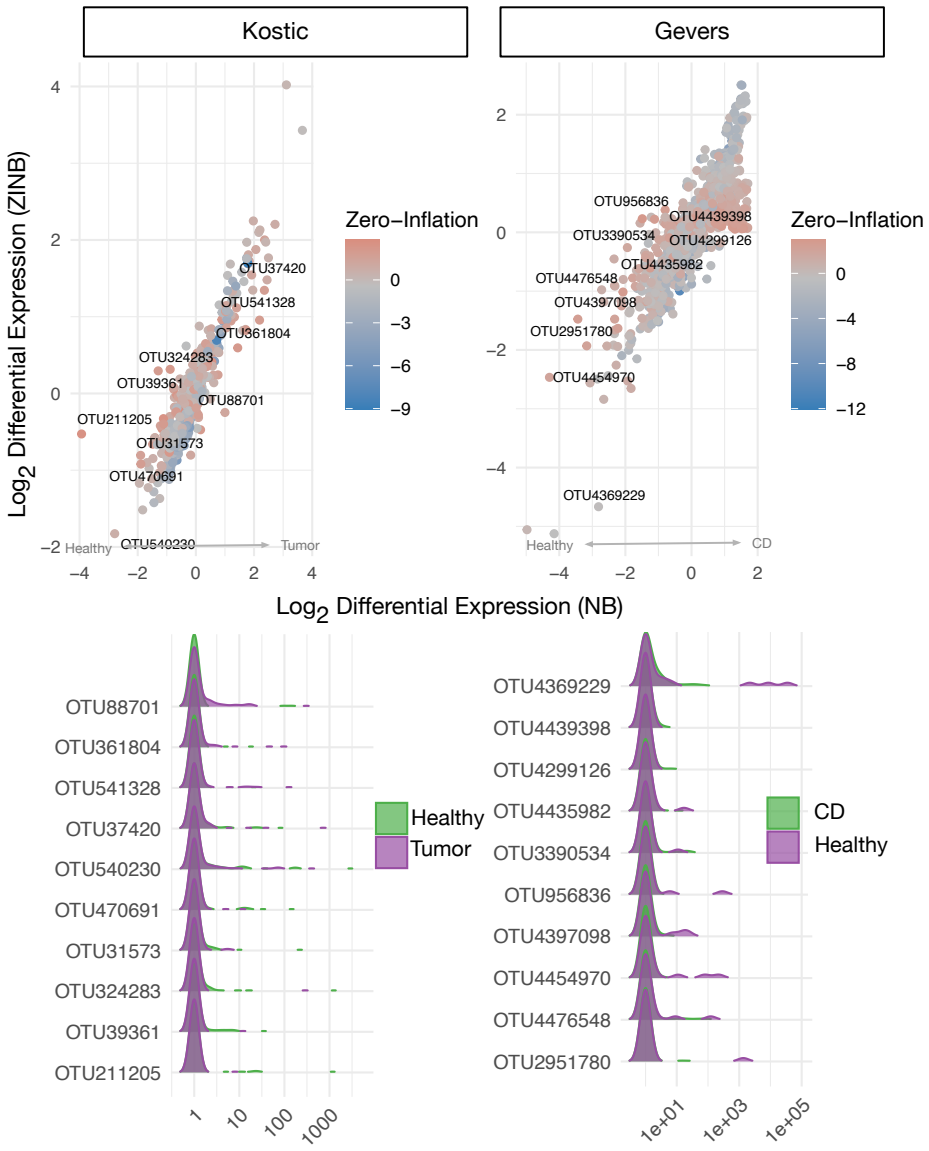

Figure S6: Differential expression (DE) estimates from a negative binomial (NB) and non-condition-specific zero-inflated negative binomial (ncsZINB) model can differ substantially. Log base 2 differential expression for the ncsZINB and NB models are shown after each was applied to two different 16S rRNA surveys. Dots represent different taxa, and each is colored according to the degree of zero-inflation as estimated by the ncsZINB 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.

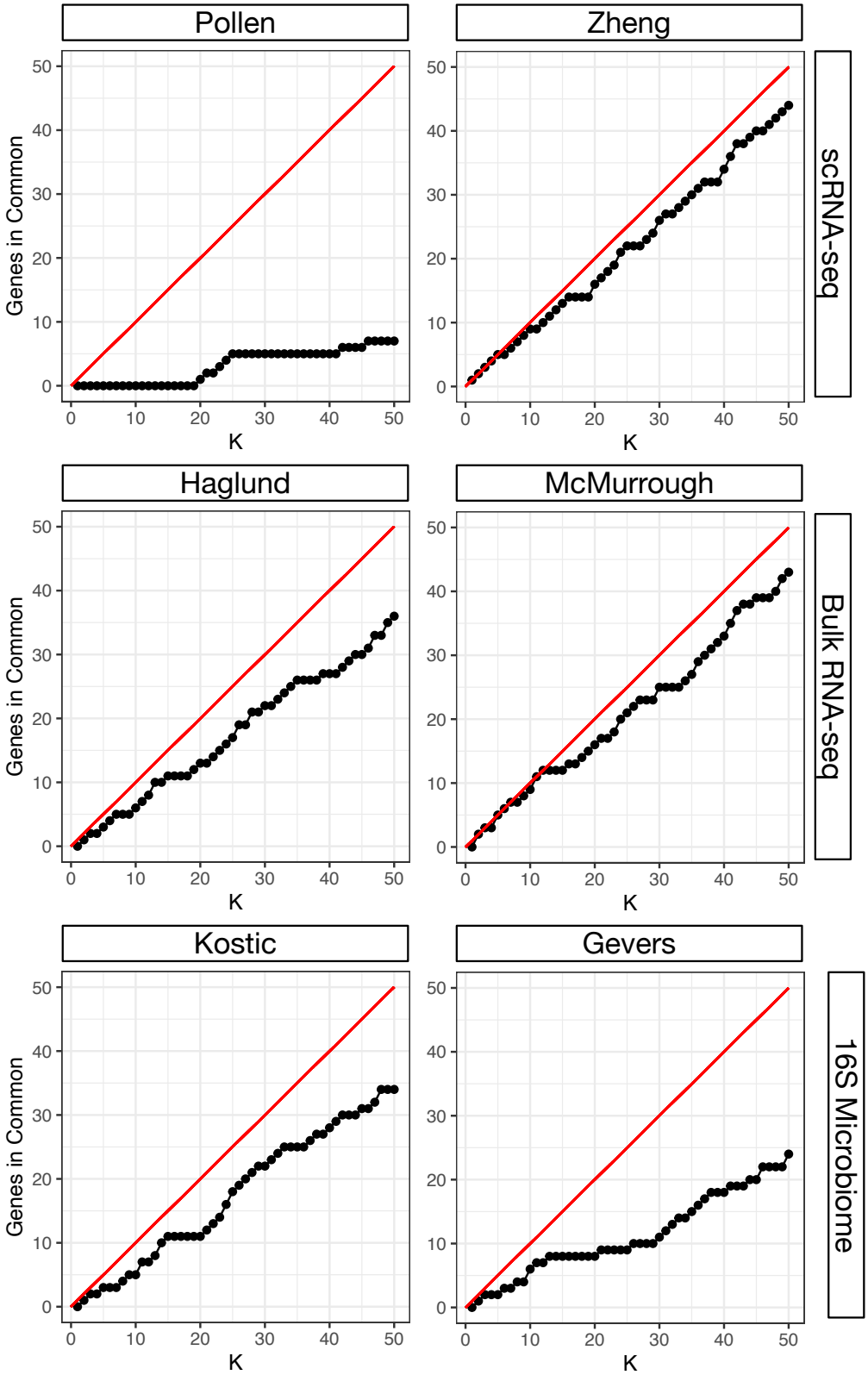

Figure S7: The ncsZINB and NB models often disagree regarding which sequences are the most differential expressed. For each dataset the intersection between the top-K most deferentially expressed sequences according to the NB and ncsZINB models is shown as a function of K.

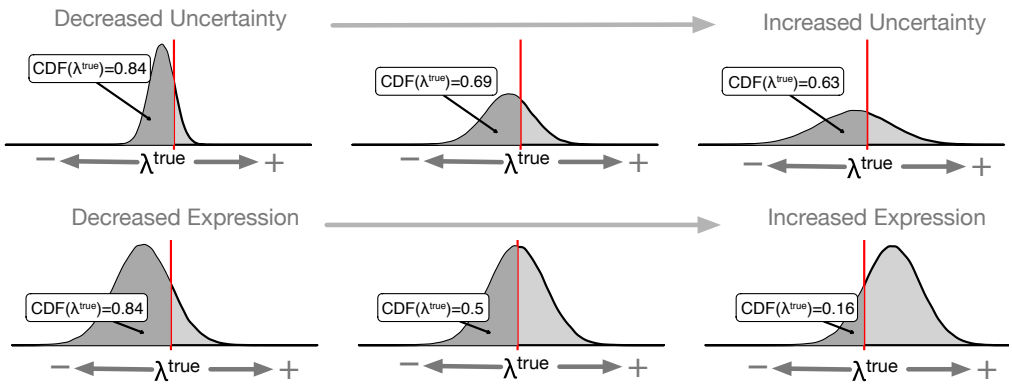

Figure S8: The cumulative distribution function (CDF) of the posterior distribution is a function that for any value of λ calculates the integral of the density from negative infinity to the specified value of λ. If we set that value of λ to be equal to the true value of λ in the simulation, then $CDF(\lambda^{true})$ is a measure of how close the mean of the density is to the true value as well as accounting for how diffuse the density is about the true value. When each density represents the posterior distribution of a model then this statistic makes a suitable performance measure for how accurately and how precisely a model inferred the true value of λ. An optimal model will produce a posterior distribution where $CDF(\lambda^{true}) = 0.5$.

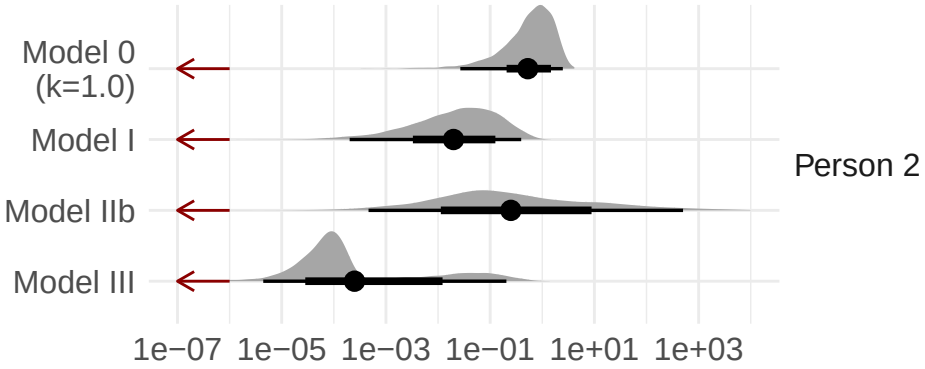

Figure S9: Posterior distribution of λ from each model applied to Simulation 5 shown on a log-scale for Person 2–an example of biological zeros and sampling zeros. Dark red arrow points to the true value of λ (negative infinity in log-space). Posterior mean as well as the 66% and 95% credible intervals are shown in black.

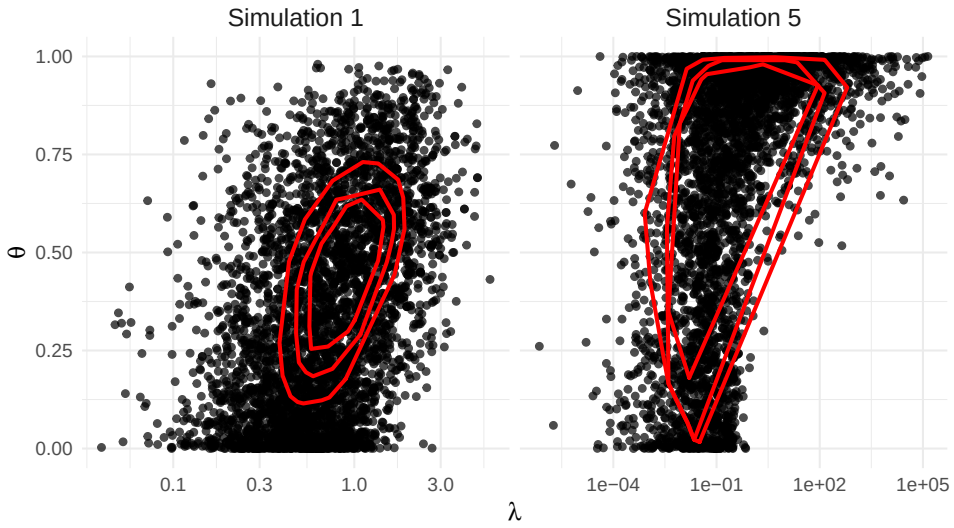

Figure S10: Large uncertainty explains the parameter inflation observed with the ZIP model. Posterior samples of λ (transcript abundance) and θ (probability of zero inflation) for the ZIP model applied to simulation 1 (sampling zeros) and simulation 5 (biological zeros). For simulation 5, the posterior distribution is of λ_2 and θ_2. The ZIP model is unable to distinguish between zeros due to sampling (i.e., low λ and low θ) versus zeros due to zero inflation (i.e., high θ and either low or high λ). Note that for Simulation 5, this uncertainty over λ_2 spans nearly 10 orders of magnitude. The 80%, 90%, and 95% highest posterior density regions for the log posterior probability are shown in red.

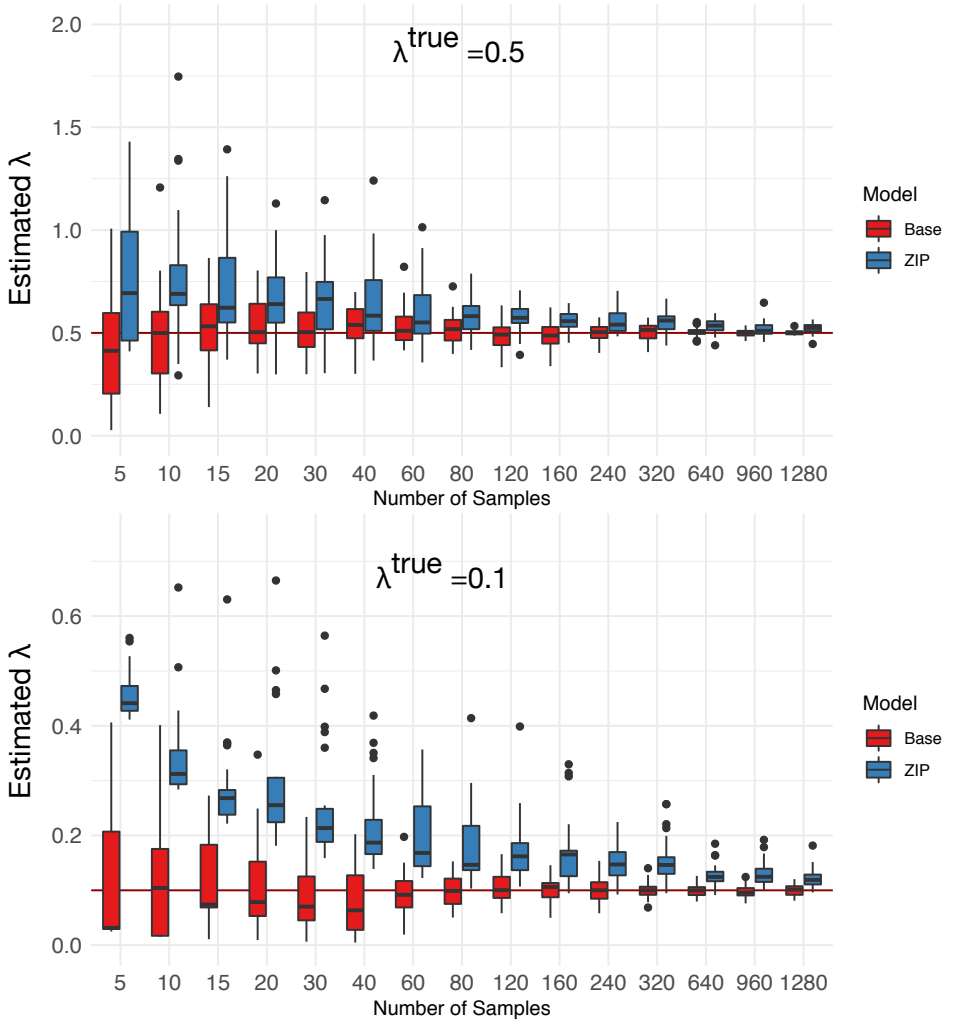

Figure S11: With sample sizes between 5 and 1280, 30 datasets were simulated and analyzed with the Base and ZIP models. Each simulated dataset contained only sampling zeros (simulated with Poisson rate (λ) parameter of 0.5 - top panel and 0.1 - bottom panel). Estimates from the ZIP model show substantial bias even for sample sizes larger than 1000. This bias increases, and takes more samples to mitigate as λ^{true} decreases.

References

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

Dataset	Sequencing Type	Grouping Variable	Compared Groups	Data Sparsity after Preprocessing	Number of Unique Sequences after Preprocessing	Other Comments	Dataset Availability
Pollen	scRNA-seq	Biological_Condition	GW21 vs. NPC	55%	9087	Coverage_Depth included as variable in model based on Vignette of Risso et al.	scRNAseq Package - dataset "fluidigm"
Zheng	scRNA-seq	group	monocyte vs. cytotoxict	58%	770	Used Assay "count_lstpm". Gene mames were shortened (zeros removed) but kept unique for making figures.	Dataset 10XMonoCytoT.rds from http://imlspenticton.uzh.ch/robinson_lab/conquer_de_comparison/
Kostic	16S Microbiome	DIAGNOSIS	Healthy vs. Tumor	72%	548	Sampes with >500 counts were retained.	Available from the Phyloseq Package as file study_1457_split_library_seqs_and_mapping.zip
Gevers	16S Microbiome	diagnosis	Healthy vs. CD	75%	1000	Only samples from the Terminal ileum were included in analysis. Only samples with >500 counts were retained. Due to computational complexity, only the 1000 most variable taxa were retained.	Dataset RISK_CCFA from R package MicrobeDS
McMurrough	Bulk RNA-seq	N/A	First 7 samples are "NS", Second 7 "S"	37%	1600		Dataset Selex available from Aldex2 pacakge
Haglund	Bulk RNA-seq	treatment	DPN vs. Control	4%	19720	The covariate time was included in the model to account for differences between samples caused by sampling time	Dataset parathyroidGenesSE in the package parathyroidSE