Biorxiv 000, 000–000 DOI: 000
000 0000

A bivariate zero-inflated negative binomial model for identifying underlying dependence with application to single cell RNA sequencing data

Hunyong Cho[1*], Chuwen Liu[1], John S. Preisser[1], and Di Wu[1,2]**

[1] *Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599,*

[2] *Department of Periodontology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599*

**email:* hunycho@live.unc.edu

***email:* did@email.unc.edu

Summary: Measuring gene-gene dependence in single cell RNA sequencing (scRNA-seq) count data is often of interest and remains challenging, because an unidentified portion of the zero counts represent non-detected RNA due to technical reasons. Conventional statistical methods that fail to account for technical zeros incorrectly measure the dependence among genes. To address this problem, we propose a bivariate zero-inflated negative binomial (BZINB) model constructed using a bivariate Poisson-gamma mixture with dropout indicators for the technical (excess) zeros. Parameters are estimated based on the EM algorithm and are used to measure the underlying dependence by decomposing the two sources of zeros. Compared to existing models, the proposed BZINB model is specifically designed for estimating dependence and is more flexible, while preserving the marginal zero-inflated negative binomial distributions. Additionally, it has a simple latent variable framework, allowing parameters to have clear and intuitive interpretations, and its computation is feasible with large scale data. Using a recent scRNA-seq dataset, we illustrate model fitting and how the model-based measures can be different from naive measures. The inferential ability of the proposed model is evaluated in a simulation study. An `R` package '`bzinb`' is available on CRAN.

Key words: Bivariate count model; Correlation; dropout; EM algorithm; Negative binomial; Single cell RNA sequencing; Zero-inflation.

1. Introduction

Single cell RNA sequencing (scRNA-seq) is a high throughput sequencing technology that profiles gene expression at a cell's resolution (Kolodziejczyk et al., 2015). This is in contrast to bulk RNA sequencing (RNA-seq), where a group of cells are sequenced altogether and consequently no cell-level information is available in data. As a price for cell-level resolution, scRNA-seq loses some information by the so-called "dropout" phenomenon; during sequencing steps (and capturing steps, e.g., in 10X sequencing platform) of scRNA-seq, a large amount of RNAs are undetected. Consequently, the observed count data include a greater number of zeros than reality (Risso et al., 2018, Hicks et al., 2017, Huang et al., 2018). That is, an expressed gene in a cell might be recorded as zero due to low transcriptome capture and sequencing efficiency (Huang et al., 2018). The artificially generated zeros due to dropouts are "technical zeros," and they are distinct from "biological" (or "real") zeros that are observed when genes were not actually expressed at the time of sample collection. In contrast, in a bulk RNA-seq, zeros mostly represent real zeros (Hicks et al., 2017).

Statistical inferences at both individual gene level (Iacono et al. (2019) and Yu (2018)) and gene set level, e.g., pathways, can be misleading without considering technical zeros. Inference of gene-gene dependence, e.g., the correlation-based method, has been widely used in pathway analysis of bulk RNA-seq data (Zhang and Horvath, 2005), and recently also used in scRNAseq data analysis (Iacono et al., 2019; Yu, 2018; Pont et al., 2019; Van Dijk et al., 2018; Eraslan et al., 2019). However, correlation between two genes with technical zeros in the scRNAseq will not reflect the true gene-gene dependence.

For example, a pair of genes, of which true expressions are highly correlated, would have less correlation, based on the observed data, when only one of them have a large amount of dropouts. On the other hand, a pair of uncorrelated genes would have higher correlation, when both genes have dropouts in a substantial portion of the sample. The systematic bias will not vanish without adjusting for the technical zeros, regardless of what dependence measure is used. This includes mutual information, $MI(X,Y) := \int\int_{\mathcal{X}\times\mathcal{Y}} f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy$ (Mc Mahon et al., 2014, Chan et al., 2017).

Two strategies have been considered to address bias generally in scRNA-seq data. Imputation methods (Li and Li, 2018 and Peng et al., 2019) aim to provide expression levels free of technical zeros by imputing some of the zeros. While imputation methods are versatile in that they provide ready-to-use data, they are not deterministic, having different results for every implementation. The second strategy is estimation of the count distribution. Once having obtained information about the distribution of true expressions and technical zeros, one can do downstream analyses such as measuring the dependence of the true expressions. Models such as SAVER (Huang et al., 2018) and DESCEND (Wang et al., 2018) have been proposed to estimate the count distribution of scRNA-seq data. For example, correlations can be calculated from SAVER-recovered genes in unique molecule index (UMI)-based DropSeq scRNA-seq data where its result is close to that measured from the “gold standard” RNA fluorescence in situ hybridization (FISH) (Huang et al., 2018). However, many of the methods taking this approach focus on modeling marginal distributions and they do not explicitly posit dependence structure between two genes.

Our proposed method takes the distribution estimation approach where a bivariate distribution explicitly addresses the dependence structure. Specifically, our method is built on a bivariate generalization of the zero-inflated negative binomial (ZINB) model. For univariate count data, zero-inflated negative binomial (ZINB) models have been well accepted and have greater capability than Poisson, zero-inflated Poisson, and negative binomial models in terms of handling augmented zeros and overdispersion. While negative binomial models have been extensively used for bulk RNA-seq data without much zero-inflation (Love et al., 2014, Robinson et al., 2010), ZINB models are typically used for scRNA-seq data (van den

Berge et al., 2018, Risso et al., 2018). Therefore, we have a particular interest in a bivariate generalization of ZINB models to model dependence of two genes in scRNA-seq data.

In consideration of underlying dependence in scRNA-seq data, it is noteworthy that there have been proposed a variety of bivariate models that fit overdispersed count data: bivariate Poisson mixture models (Gurmu and Elder, 1999, Famoye, 2010, and Jørgensen, 1987), bivariate generalized Poisson models (Famoye and Consul, 1995) and copula models (Cameron et al., 2004). These models can be further extended to flexibly accommodate excess zeros by introducing zero-inflation parameters or composing hurdle models. For a comprehensive survey of bivariate count models, refer to Cameron and Trivedi (2013) and Chou and Steenhard (2011).

Of a plethora of the proposed models in the literature, many of the bivariate Poisson mixture models and bivariate generalized Poisson models take overly complicated forms, they do not have simple marginal distributions (e.g., GBIVARNB model in Gurmu and Elder, 1999), and their parameters are hard to interpret and/or computationally expensive to estimate. Copula-based bivariate models can be alternatives to the mixture models, but they depend on the underlying copula models and can be difficult to interpret.

Many existing bivariate negative binomial models are primarily designed for modeling marginal means rather than pairwise dependence. For example, Gurmu and Elder (1999) discussed a bivariate negative binomial distribution (BIVARNB), but their model is specified by only four parameters, which may not provide sufficient flexibility to delineate diverse distributional structure. For such a bivariate joint distribution, four parameters are needed to specify the first two marginal moments of each of the two independent variables, while another parameter is needed solely for modeling the dependence. Subsequently Wang (2003) extended BIVARNB to a zero-inflated BIVARNB regression setting. In this model, zero-inflation is dictated by a single parameter, implying that when one variable either drops out or not, the other variable behaves exactly the same, which may not be the case for scRNA-seq data; one gene can drop out, while the other does not. Instead, it is possible to have three free parameters for the full joint zero-inflation probability structure (Li et al., 1999).

We propose a bivariate zero-inflated negative binomial model with eight parameters: five parameters for the negative binomial part and another three free parameters for the zero-inflation part. This model allows analyzing the dependence of two zero-inflated count variables parametrically but with more flexibility than existing models. That is, the five parameters of our proposed model can characterize all the five moments of the first two orders, and the three zero-inflation parameters can model the dropouts with full flexibility.

The rest of the paper is organized as follows. In Section 2, we describe how the model is constructed in the order of a Bivariate Negative Binomial model and a Bivariate Zero-inflated Negative Binomial model. We present the maximum likelihood estimator using the expectation-maximization (EM) algorithm in Section 3. In Section 4, we illustrate how well the models fit data and how model-based dependence measures behave in contrast to naive measures using real scRNA-seq data. Then in Section 5, we show how point and interval estimators perform based on simulations. In Section 6, we address limitations of the models and discuss potential extensions. Section 7 provides software information.

2. The model

2.1 *A Bivariate Negative Binomial Model*

In constructing the BZINB model, to induce dependence and zero-inflation, layers of latent variables were used as in Kocherlakota and Kocherlakota (1992) and Li et al. (1999). We first introduce a simpler model, the Bivariate Negative Binomial (BNB) model in this subsection, and then generalize it to Bivariate Zero-Inflated Negative Binomial (BZINB) model in Subsection 2.2.

One of the key assumptions about the dependence structure of BNB (and BZINB) is that the mean parameters of two Poisson random variables are gamma random variables that share a common gamma random variable. Let $R_j \sim Gamma(\alpha_j, \beta)$ for $j = 0, 1, 2$, where α_j and β are the shape and scale parameters, respectively. Then $(R_0 + R_1, R_0 + R_2)$ is bivariate gamma distributed, denoted as $BGamma(\alpha_0, \alpha_1, \alpha_2, \beta)$. To account for heterogeneous scales of the two Poisson mean variables, we introduce an additional parameter $\delta \in R^+$. Then, a pair (X_1, X_2) of Poisson variables with means $(R_0 + R_1, \delta(R_0 + R_2))$ follow a bivariate negative binomial distribution, denoted as

$$(X_1, X_2) \sim BNB(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2), \tag{1}$$

where we reparametrize (β, δ) as $(\beta_1, \beta_2) = (\beta, \delta\beta)$ and the observed density is given as,

$$\begin{aligned}
&f_{BNB}(x_1, x_2) \\
&= \iiint\limits_{R_+^3} \frac{(R_0+R_1)^{x_1}(R_0+R_2)^{x_2} e^{-\frac{1+\beta_1+\beta_2}{\beta_1}R_0 - \frac{1+\beta_1}{\beta_1}R_1 - \frac{1+\beta_2}{\beta_1}R_2} R_0^{\alpha_0-1} R_1^{\alpha_1-1} R_2^{\alpha_2-1} \beta_2^{x_2}}{x_1! x_2! \Gamma(\alpha_0)\Gamma(\alpha_1)\Gamma(\alpha_2)\beta_1^{\alpha_0+\alpha_1+\alpha_2+x_2}} \prod_{j=0}^{2} dR_j \\
&\quad \times 1_{(x_1,x_2)\in N_0^2} \\
&= \sum_{k=0}^{x_1}\sum_{m=0}^{x_2} \binom{\alpha_0 + x_1 + x_2 - k - m - 1}{\alpha_0 + x_2 - m - 1} \binom{\alpha_0 + x_2 - m - 1}{\alpha_0 - 1} \binom{\alpha_1 + k - 1}{\alpha_1 - 1} \binom{\alpha_2 + m - 1}{\alpha_2 - 1} \\
&\quad \times \frac{\beta_1^{x_1}\beta_2^{x_2}(\beta_1+\beta_2+1)^{k+m-x_1-x_2-\alpha_0}}{(\beta_1+1)^{k+\alpha_1}(\beta_2+1)^{m+\alpha_2}} 1_{(x_1,x_2)\in N_0^2},
\end{aligned}$$

where N_0 denotes the nonnegative integer space, and superscripts represent the dimension of the product space. The support indicators will be omitted throughout this paper when the context is clear.

This bivariate negative binomial model (BNB) is marginally negative binomial, as we know from the construction procedure that both X_1 and X_2 are Poisson random variables with means marginally Gamma distributed, respectively:

$$X_j \sim NB(\alpha_0 + \alpha_j, \frac{1}{\beta_j + 1}) \text{ for } j = 1, 2,$$

where the random variable $X \sim NB(\nu, \phi)$ can be interpreted as the minimum number of failures to have ν successes with probability of ϕ for each trial; i.e., its density is expressed as $f_{NB}(x;\nu,\phi) = \binom{x+\nu-1}{x}\phi^{\nu}(1-\phi)^{x}$.

Interpretation of the BNB parameters is straightforward: α_0, α_1, and α_2 are the shape parameters of latent variables, where the larger α_0 implies a larger amount of shared components in X_1 and X_2 and thus larger correlation; β_1 and β_2 controls the scale of X_1 and X_2, respectively. Note in scRNA-seq data context, X_1 and X_2 may represent the *true* expression level of each of two genes in a cell in the absence of dropout events, which we rarely observe in practice.

The first two moments and the correlation of a BNB random pair are given as,

$$\begin{aligned}
E(X_j) &= (\alpha_0+\alpha_j)\beta_j && j=1,2\\
Var(X_j) &= (\alpha_0+\alpha_j)\beta_j(\beta_j+1) && j=1,2\\
Cov(X_1,X_2) &= \alpha_0\beta_1\beta_2\\
Cor(X_1,X_2) &= \frac{\alpha_0}{\sqrt{(\alpha_0+\alpha_1)(\alpha_0+\alpha_2)}}\sqrt{\frac{\beta_1\beta_2}{(\beta_1+1)(\beta_2+1)}} && (2)
\end{aligned}$$

Note that this distribution only allows positive correlation. See Section 6 for more discussion.

Maher (1990) developed another bivariate negative binomial distribution that is a constrained case of BNB in a sense that the marginal means and variances are the same for both variables.

One can further generalize this BNB model into a m-variate negative binomial model by adding common latent gamma parameter(s) to the m gamma variables.

2.2 *A Bivariate Zero-inflated Negative Binomial Model*

In this subsection, we generalize BNB model to BZINB model by including zero-inflation components. Since BZINB is also a generalization of univariate zero-inflated negative binomial model (ZINB), we illustrate the construction of univariate ZINB model first and move to the bivariate version.

A univariate negative binomial model, $NB(\nu, \phi)$, can be generalized to allow zero-inflation by having an additional parameter, π: $ZINB(\nu, \phi, \pi)$. The zero-inflated negative binomial (ZINB) model has a latent variable interpretation. Let X follow $NB(\nu, \phi)$ and E denote the zero-inflation indicator having 1 with probability of π and 0 otherwise, independently of X. Then $Y \equiv (1-E)X$ follows $ZINB(\nu, \phi, \pi)$ with the density of $f_{ZINB}(y; \nu, \phi, \pi) = (1-\pi) f_{NB}(y; \nu, \phi) + \pi\zeta(y)$, where $\zeta(a) \equiv 1_{(a=0)}$.

Similarly, a multivariate zero-inflated random variable can be constructed using a latent variable that follows the multivariate Bernoulli distribution as in the Poisson case (Li et al., 1999). For a bivariate distribution, suppose we have a random vector $\boldsymbol{E} \equiv (E_1, E_2, E_3, E_4)^\top \sim MN(1, \boldsymbol{\pi})$, where $MN(1, \boldsymbol{\pi})$ denotes the multinomial distribution with a single trial and an associated probability of $\boldsymbol{\pi} \equiv (\pi_1, \pi_2, \pi_3, \pi_4)^\top$. Now the bivariate zero-inflated negative binomial distribution (BZINB) can be formulated as:

$$(Y_1, Y_2) := ((E_1 + E_2)X_1, (E_1 + E_3)X_2), \tag{3}$$

where $(X_1, X_2) \sim BNB(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2)$ and E_1, E_2, E_3 and E_4 are the indicators of observing both X_1 and X_2, only X_1, only X_2, and none of them, respectively. We say $(Y_1, Y_2) \sim BZINB(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2, \pi_1, \pi_2, \pi_3, \pi_4)$. A simpler model with a restriction of $\pi_2 = \pi_3 = 0$ can also be considered as in Wang (2003).

The density of a BZINB variable is

$$\begin{aligned} & f_{BZINB}(y_1, y_2; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}) \\ &= \pi_1 f_{BNB}(y_1, y_2; \alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2) + \pi_2 f_{NB}(y_1; \alpha_0 + \alpha_1, \frac{1}{\beta_1 + 1})\zeta(y_2) \\ &\quad + \pi_3 f_{NB}(y_2; \alpha_0 + \alpha_2, \frac{1}{\beta_2 + 1})\zeta(y_1) + \pi_4 \zeta(y_1 + y_2), \end{aligned}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)^\top, \boldsymbol{\beta} = (\beta_1, \beta_2)^\top$, and $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)^\top$ with $\mathbf{1}^\top \boldsymbol{\pi} = 1$.

Here, the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have the same interpretation as in BNB but in the presence of dropouts, and $\boldsymbol{\pi}$ indicates the dropout probability, where π_1, π_2, π_3, and π_4 are the probability that none, Y_2 only, Y_1 only, and both were dropped out, respectively.

In scRNA-seq data, Y_1 and Y_2 are the *recorded* number of expressions for each of two genes in a cell. The term *recorded* was used in contrast to *true* in a sense that an unobserved subset of the zeros are technical zeros due to dropouts.

This BZINB distribution is marginally ZINB, since the latent random variables, X_1 and X_2, are marginally negative binomial random variables (from Subsection 2.1) with probabilities of being observed, $\pi_1 + \pi_2$ and $\pi_1 + \pi_3$, respectively:

$$Y_j \sim ZINB(\alpha_0 + \alpha_j, \frac{1}{\beta_j + 1}, \pi_{4-j} + \pi_4) \text{ for } j = 1, 2. \tag{4}$$

The first two moments of a BZINB pair are given as,

$$\begin{aligned}
E(Y_j) &= (\pi_1 + \pi_{j+1})(\alpha_0 + \alpha_j)\beta_j & j = 1, 2 \\
Var(Y_j) &= (\alpha_0 + \alpha_j)^2 \beta_j^2 (\pi_1 + \pi_{j+1})(1 - \pi_1 - \pi_{j+1}) & \\
&\quad + (\alpha_0 + \alpha_j)\beta_j(\beta_j + 1)(\pi_1 + \pi_{j+1}) & j = 1, 2 \\
Cov(Y_1, Y_2) &= \{\alpha_0 + (\alpha_0 + \alpha_1)(\alpha_0 + \alpha_2)\}\beta_1\beta_2\pi_1 & \\
&\quad - (\alpha_0 + \alpha_1)(\alpha_0 + \alpha_2)\beta_1\beta_2(\pi_1 + \pi_2)(\pi_1 + \pi_3), &
\end{aligned}$$

and the correlation $\rho(Y_1, Y_2)$ is not further simplified than $Cov(Y_1, Y_2)/\sqrt{Var(Y_1)Var(Y_2)}$.

When dropouts are not real zeros but instead represent non-zero counts caused by technical reasons, then the true underlying correlation ρ^* of Y_1 and Y_2 under BZINB model is simply the correlation of X_1 and X_2 (Equation (2)), which is

$$\rho^*(Y_1, Y_2) = \frac{\alpha_0}{\sqrt{(\alpha_0 + \alpha_1)(\alpha_0 + \alpha_2)}} \sqrt{\frac{\beta_1\beta_2}{(\beta_1 + 1)(\beta_2 + 1)}}. \tag{5}$$

3. Estimation

With the natural interpretation of BZINB model as layers of latent variables, one can estimate the parameters by the expectation-maximization (EM) algorithm.

The complete density is given as,

$$f(Y_1, Y_2, X_1, X_2, R_0, R_1, R_2, E_1, E_2, E_3, E_4)$$

$$= f(X_1, X_2, R_0, R_1, R_2, E_1, E_2, E_3, E_4) \times 1_{(Y_1 = X_1(E_1+E_2), Y_2 = X_2(E_1+E_3))}$$

with

$$f(X_1, X_2, R_0, R_1, R_2, E_1, E_2, E_3, E_4)$$

$$= \frac{(R_0+R_1)^{X_1}(R_0+R_2)^{X_2} R_0^{\alpha_0-1} R_1^{\alpha_1-1} R_2^{\alpha_2-1} \beta_2^{X_2} \prod_{k=1}^4 \pi_k^{E_k}}{X_1! X_2! \Gamma(\alpha_0)\Gamma(\alpha_1)\Gamma(\alpha_2) \exp\{R_0 \frac{1+\beta_1+\beta_2}{\beta_1} + R_1 \frac{1+\beta_1}{\beta_1} + R_2 \frac{1+\beta_2}{\beta_1}\} \beta_1^{X_2+\alpha_0+\alpha_1+\alpha_2}}$$

$$\times 1_{\sum_{k=1}^4 E_k = 1}.$$

Thus, the full individual log-likelihood for the ith entry, or the ith cell, is

$$l_i^{\text{Full}}$$

$$= X_{1,i} \log(R_{0,i} + R_{1,i}) + X_{2,i} \log(R_{0,i} + R_{2,i})$$

$$+ (\alpha_0 - 1)\log R_{0,i} + (\alpha_1 - 1)\log R_{1,i} + (\alpha_2 - 1)\log R_{2,i}$$

$$+ X_{2,i}\log\beta_2 - (X_{2,i} + \alpha_0 + \alpha_1 + \alpha_2)\log\beta_1 + \sum_{k=1}^{4} E_{k,i}\log\pi_k - \log X_{1,i}! - \log X_{2,i}!$$

$$- \log\Gamma(\alpha_0) - \log\Gamma(\alpha_1) - \log\Gamma(\alpha_2) - R_{0,i}\frac{1+\beta_1+\beta_2}{\beta_1} - R_{1,i}\frac{1+\beta_1}{\beta_1} - R_{2,i}\frac{1+\beta_2}{\beta_1}$$

$$+ \log 1_{(Y_{1,i} = X_{1,i}(E_{1,i}+E_{2,i}))} + \log 1_{(Y_{2,i} = X_{2,i}(E_{1,i}+E_{3,i}))} + \log 1_{\sum_{k=1}^4 E_k = 1}.$$

The expected full log-likelihood conditional on the observed data is linear in $E[R_{j,i}|Y_{1,i}, Y_{2,i}; \boldsymbol{\theta}]$, $E[\log(R_{j,i}|Y_{1,i}, Y_{2,i}; \boldsymbol{\theta})]$, $E[E_{k,i}|Y_{1,i}, Y_{2,i}; \boldsymbol{\theta}]$, and $E[X_{2,i}|Y_{1,i}, Y_{2,i}; \boldsymbol{\theta}]$, where $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\pi}^\top)^\top$, $j = 0, 1, 2$ and $k = 1, 2, 3, 4$. The formulae of the components are given in Web Appendix A.

As the likelihood is the product of functions convex with respect to each of the parameters, the maximization can be achieved by solving a system of score equations. The individual scores are given as:

$$\begin{aligned}
\partial_{\alpha_j} E[l_i^{Full}|\cdot] &= E[\log R_{j,i}|\cdot] - \log\beta_1 - \psi(\alpha_j) && j = 0,1,2\\
\partial_{\beta_1} E[l_i^{Full}|\cdot] &= E[R_{0,i}+R_{2,i}|\cdot]\frac{1+\beta_2}{\beta_1^2} + \frac{E[R_{1,i}|\cdot]}{\beta_1^2} - \frac{\alpha_0+\alpha_1+\alpha_2+E[X_{2,i}|\cdot]}{\beta_1}\\
\partial_{\beta_2} E[l_i^{Full}|\cdot] &= -\frac{E[R_{0,i}+R_{2,i}|\cdot]}{\beta_1} + \frac{E[X_{2,i}|\cdot]}{\beta_2}\\
\partial_{\pi_j} E[l_i^{Full}|\cdot] &= \frac{E[E_{j,i}|\cdot]}{\pi_j} - \frac{1-E[E_{j,i}|\cdot]}{1-\pi_j} && j = 1,2,3,
\end{aligned}$$

where the conditioning arguments $(\boldsymbol{Y}_1, \boldsymbol{Y}_2; \boldsymbol{\theta})$ are suppressed as $(\cdot)$ and can be replaced with $(Y_{1,i}, Y_{2,i}; \boldsymbol{\theta})$ where we assume a sample of independent entries, $\boldsymbol{Y}_l$ denotes $(Y_{l,1}, ..., Y_{l,n})^\top$ for $l = 1, 2$, n is the sample size, and $\partial_a b$ denotes the partial derivative of b with respect to a.

At the $k+1$st iteration of the EM algorithm, we get $\boldsymbol{\theta}^{(k+1)}$ by solving the score equations $\partial_{\boldsymbol{\theta}} \sum_i^n E[l_i^{Full}|\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{\theta}^{(k)}] = \mathbf{0}$:

$$\begin{aligned}
\frac{\beta_2^{(k+1)}}{\beta_1^{(k+1)}} &= \frac{\bar{E}[X_{2,i}|\cdot]}{\bar{E}[R_{0,i}+R_{2,i}|\cdot]}\\
\beta_1^{(k+1)} &= \frac{\bar{E}[R_{0,i}+R_{1,i}+R_{2,i}|\cdot]}{\alpha_0^{(k+1)}+\alpha_1^{(k+1)}+\alpha_2^{(k+1)}}\\
\pi_j^{(k+1)} &= \bar{E}[E_{j,i}|\cdot] && j = 1,2,3,4\\
\alpha_j^{(k+1)} &= \psi^{-1}\{-\log\beta_1^{(k+1)} + \bar{E}[\log R_{j,i}|\cdot]\} && j = 0,1,2,
\end{aligned}$$

where $\bar{E}[A|\cdot]$ denotes the empirical average of the conditional expectations, i.e., $\frac{1}{n}\sum_i^n E[A_i|\cdot]$, $\psi(\cdot)$ is the digamma function, and the conditioning arguments $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{\theta}^{(k)})$ are again suppressed. The equations can be solved by solving the following through Newton-Raphson algorithm:

$$\text{Solve for } \beta_1 = \frac{\bar{E}[R_0+R_1+R_2|\cdot]}{\sum_{k=0}^2 \psi^{-1}(-\log\beta_1 + \bar{E}[\log R_k|\cdot])}.$$

$$\text{Then get } \alpha_j = \psi^{-1}(-\log\beta_1 + \bar{E}[\log R_j|\cdot]).$$

After iterations enough to observe convergence, the final updated parameter values serve as the maximum likelihood estimate.

The standard error of the maximum likelihood parameter estimates can be calculated using observed information. In Web Appendix B, detailed formulae are given, and simulations illustrating the accuracy of standard error estimation are included in Section 5.

4. Model and measure comparisons based on real data

4.1 *Model comparison using real data*

In this section, we show how the BZINB model fits real scRNA-seq data compared to its nested models (in Subsection 4.1) and present how model-based dependence measures can be different from naive measures (in Subsection 4.2). The data were collected from paneth cells of C57Bl6 mouse with a Sox9 gene knockout. The Fluidigm C1 system was used to capture single cells and generate Illumina libraries using manufacutrers' protocols. Illumina NextSeq sequencing platform was used for paired end sequencing. Reads per cell were demultiplexed using mRNASeqHT_demultiplex.pl, a script provided by Fluidigm. Low quality base calls and primers were removed using Trimmomatic (Bolger et al., 2014) and poly-A tails were removed using a custom perl script. Reads were aligned to the mouse genome (mm9) using STAR (https://academic.oup.com/bioinformatics/article/29/1/15/272537) and read per gene were counted using htseq-count (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4287950/). The data are composed of 23,425 genes for 800 cells, where all the cells came from a single mouse and have the same cell type. Over 90% of genes have more than 90% zero counts and the average proportion of zero counts for a gene is 97.3% in these zero-inflated data.

We compare four nested models: BZINB, BNB, bivariate zero-inflated Poisson (BZIP), and bivariate Poisson (BP). BZIP has fixed mean values instead of latent gamma variables of BZINB, and BP further lacks zero-inflation components. The estimated densities of these models are compared with the empirical density for 50 gene pairs.

To systematically study the model performances, we performed stratified sampling of genes according to their proportion of zeros; strata H, M, L, and V include genes with $\geqslant 90\%$, 80% to 90%, 60% to 80%, and $< 60\%$ zeros, respectively. Genes with $\geqslant 98\%$ of zeros and genes with extremely large expression ($> 10,000$ counts for at least one cell) were screened out. After screening out those irregular genes, each group has 81.4%, 13.5%, 4.2%, and 0.9% of genes in the order.

We randomly selected 5 pairs from each possible combination of two strata (HH, MM, LL, VV, HM, HL, HV, ML, MV, and LV) without replacement. For each of the 50 pairs (5 pairs × 10 combinations), we estimated the parameters of the four nested models. Based on the parameter estimates, the distributions of the four models were compared. As it is not straightforward to compare the estimated model-based densities with the empirical density, we drew a random sample of size $n = 800$ from each estimated model and the resulting empirical densities were then compared (Figure 1 for several pairs and Web Figure 1 for all the 50 pairs). As we cannot preclude the chance of getting unlikely instances by doing Monte Carlo sampling, we added results of two more replicates in Web Figures 2 and 3. We furthermore illustrate the exact values of the estimated density in Figure 2 for a couple of pairs and in Web Figure 4 for all the pairs.

Figure 1 illustrates the real and the model-based empirical distributions for the first pairs of 10 combinations. The results including all 50 pairs and their replicates can be found in Web Figures 1 to 3. For any pair, the BP model obviously fails to address the overdispersion and zero-inflation, while the BZIP model could not properly mimic the overdispersion. BNB and BZINB seem to fairly mimic the real distribution in most of the pairs. The poor performances of Poisson-based models and decently good performances of BNB and BZINB models can also be seen on Figure 2.

However, when genes have some large-valued counts and many zeros at the same time either marginally or jointly, BZINB has an apparent advantage over BNB model. Often, in

BNB model, nonzero count pairs are highly concentrated on the diagonal line, while nonzero counts in BZINB model are more dispersed away from the diagonal line (LL1 in Figure 1 and more examples in Web Figures 1 to 3). This can be explained by the lack of flexibility of BNB model. When data are highly zero-inflated but overdispersed at the same time, BNB is forced to have small shape parameters ($\alpha_j,\ j = 0, 1, 2$) and large scale parameters ($\beta_j,\ j = 1, 2$) while keeping the mean of the latent Gamma variables, $E[R_j] = \alpha_j \beta_1$, close to zero. These latent Gamma variables, serving as mean parameters of Poisson variables, take on very small values most of the times and very large values with small chance. It is unlikely that both R_1 and R_2 have large numbers at the same time (CASE 1), but it is more frequent that R_0 alone has a large number (CASE 2). Thus, the latent Poisson variables, X_1 and X_2, are more likely to have similarly large numbers (resulting from CASE 2) than to have significantly different nonzero numbers (resulting from CASE 1).

[Figure 1 about here.]

[Figure 2 about here.]

4.2 *Real data example of dependence measures*

When the excess zeros are believed to come from dropouts, BZINB model may uncover the true underlying dependence using measures such as ρ^* and MI^*. Note that MI^* is defined similarly to that of ρ^* and can be estimated by first estimating the BZINB model parameters and by measuring the mutual information of the estimated distribution after replacing π with $(1, 0, 0, 0)^\top$.

For the same 50 pairs in the previous subsection, we estimated the dependence using naive measures – Pearson correlation (PC) and empirical mutual information (EMI) – and zero-inflation adjusted measures – underlying correlation (ρ^*) and underlying MI (MI^*) based on BZINB model. Figure 3 summarizes the estimates for all the pairs. The plots of empirical distribution with estimated dependence measures for each pair are also available on Web Figure 5.

In Figrue 3 LEFT, we see that PC and ρ^* mostly behave in the same direction, but also that they can have values in the opposite directions (e.g., HL5 and HL4). If we judge whether two genes are correlated based on (naive) Pearson correlation (PC) with a certain threshold, say PC > 0.2, many genes might be missed (e.g., HL5) or falsely included (e.g., HL4).

Similar analyses can be done for MI-based measures. Both EMI and MI^* estimates are correlated, however, there are pairs that are located away from the tendency. For example the pair MV1 has highest MI^*, while its EMI is not one of the highest. Also note that the values of MI^* are in general less than those of EMI for scRNA-seq data. Heavy proportion of zero-zero pairs boosts naive EMI, while MI^* removes the effects of the co-zero-inflation.

These results suggest that measures that fail to distinguish between technical and real zeros may be highly misleading.

[Figure 3 about here.]

5. Evaluation of estimators based on simulation

We ran simulations to study the performance of estimators of underlying correlation and the associated standard error under finite sample size. We considered 40 distinct sets of BZINB parameter values (Table 1). Note that for each of ρ^*s there are two distinct sets of parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, the first (a) of which have lower $\boldsymbol{\alpha}$ values and the second (b) of which have higher $\boldsymbol{\alpha}$ values. For each parameter set $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi})$ and for $n = 250, 500, 800, 1500, 2500$, we generated random BZINB samples of size n, $n_{\text{sim}} = 1,000$ times.

[Table 1 about here.]

For each k of n_{sim} simulation replicates, we got an estimate $\hat{\rho}^*_k$ of the parameter ρ^*, the standard error estimate $se(\hat{\rho}^*_k)$, and the logit-transformed 95% confidence interval (i.e.,

$\text{logit}^{-1}(\text{logit}(\hat{\rho}_k) \pm 1.96\frac{se(\hat{\rho}_k)}{\hat{\rho}_i(1-\hat{\rho}_k)})$). Then for each set of parameters, the following three quantities were calculated:

- the average standard error (SE, $\bar{se}(\hat{\rho}^*)$)
- the standard deviation of the parameter estimates (SD, $sd(\hat{\rho}^*)$)
- the empirical coverage probability (CP, $\frac{1}{n_{\text{sim}}}\sum_{k=1}^{n_{\text{sim}}} 1_{\rho^* \in \text{CI}_k}$, where CI_k is the logit-transformed 95% confidence interval for the kth replicate).

[Figure 4 about here.]

[Figure 5 about here.]

The simulation results are provided in Figures 4 and 5. First, the mean parameter estimates are close, or getting closer as sample size grows, to their true parameter values for each of the 40 scenarios. For most of the 40 parameter sets, CP was close to 0.95, and for those not close, CP gets closer to 0.95 with increasing sample size. In the same context, the average standard error (SE) was close to the standard deviation of the parameter estimates (SD) especially when the sample size was large. However, when the true underlying correlation was close to zero (i.e., 0.01 in our example), standard error estimation did not perform as well in terms of both CP and closeness of SE to SD. The parameter value being near the boundary may be responsible for the poorer performance.

6. Discussion

In this paper, we proposed the BZINB model that enables accurate estimation of pairwise dependence between two genes in scRNA-seq data. It models bivariate count data with high flexibility by having eight free parameters and at the same time with simple latent variable interpretations. By decomposing two sources of zeros, the distribution of counts without zero-inflation is recovered and the dependence is measured accordingly.

In our BZINB model, we assume an independent and identically distributed random bivarate sample of zero-inflated counts. One can generalize this homogeneous mean model to allow for subgroup analysis or joint conditional mean analysis by introducing the generalized linear model framework. As in the univariate ZINB regression, the latent count variables (i.e., X_1 and X_2) can be modeled using linear predictors with some link function.

Our model can be applied to other settings where there is a belief in two sources of zeros such as frailty, e.g., the first source corresponds to a cohort of people who are not susceptible to disease and will always have a zero count; the other source are random zeros among susceptible individuals. In this case, the dependence measure proposed in this article applies to the bivariate outcome among the latent class of individuals that are susceptible to disease.

When the mRNAs are perfectly captured and sequenced in all cells, zeros always indicate that genes are not expressed, i.e., they are real zeros. In these settings where the excess zeros are not caused by dropout, the overall mean count and the proportion of subjects with positive counts have meaningful interpretations that may be directly modeled by marginalized ZINB (Preisser et al., 2016) and hurdle models (Mullahy, 1986), respectively. Directly modeling the observed counts (i.e., (Y1, Y2)) using such models extended to bivariate counts could be beneficial. These scenarios underscore that any model, including BZINB, may not be ideal for all purposes, and that the statistical model for zero-inflated counts should be chosen to match the research question (Preisser et al., 2017).

In the BZINB model, allowing only positive ρ^* can be regarded as a limitation. One justification for the BZINB model is that the negative correlation of count data are not so prevalent in reality. For example, in genomics data, there are some genes that suppress other genes from being expressed, however, such genes either are relatively rare or have weak negative correlation with other genes. On the other hand, when we believe that the zeros are mostly true zeros, we can consider using the original correlation ($\rho(Y_1, Y_2)$) which allows for negative correlation, instead of $\rho^*(Y_1, Y_2)$.

An alternative to this fully parametric approach is to use the weighted Pearson correlation based on the parameter estimates. One way is to use the estimated conditional probability of no dropout as a weight for each entry. Since the estimated dropout probability is used in a form of weight, this correlation measure is more robust to model misspecification. It also allows negative correlation values.

As discussed before, the BZINB model can also be generalized to a multivariate zero-inflated negative binomial model. This model may have an exponentially increasing number of latent variables or parameters as the dimension gets large. Though the lack of parsimony may make the multivariate model look less attractive, the idea can be very practically used in simulating multivariate zero-inflated count data and potentially in statistical analysis based on Bayesian models. For instance, a genomic count data with large amount of zeros can be mimicked by a set of latent random layers along with the generalized linear model framework.

7. Software

An `R` package `bzinb` estimating BZINB parameters using EM algorithm was written in `R` version 3.5.1 (R Core Team, 2019), and is available on CRAN. The R codes for the real data example analysis and simulation study are included in Web Appendix C.

Acknowledgements

The authors thank Drs. Scott Magness and Joshua Starmer for providing the mouse paneth scRNA seq data and Dr. Terence P. Speed for helpful discussion.

References

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* **30,** 2114–2120.

Cameron, A. C., Li, T., Trivedi, P. K., and Zimmer, D. M. (2004). Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *The Econometrics Journal* **7(2),** 566–584.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data.* Cambridge university press.

Chan, T. E., Stumpf, M. P., and Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems* **5(3),** 251–267.

Chou, N. T. and Steenhard, D. (2011). Bivariate count data regression models - a SAS® macro program. Sas global forum - statistics and data analysis, SAS Institute.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications* **10,** 390.

Famoye, F. (2010). On the bivariate negative binomial regression model. *Journal of Applied Statistics* **37(6),** 969–981.

Famoye, F. and Consul, P. C. (1995). Bivariate generalized poisson distribution with some applications. *Metrika* **42(1),** 127–138.

Gurmu, S. and Elder, J. (1999). Generalized bivariate count data regression models. *Economics Letters* **68(1),** 31–36.

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2017). Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics* **19(4),** 562–578.

Huang, M., Wang, J., nd H. Dueck, E. T., Shaffer, S., Bonasio, R., ..., and Zhang, N. R. (2018). Saver: gene expression recovery for single-cell rna sequencing. *Nature methods* **15(7),** 539–542.

Iacono, G., Massoni-Badosa, R., and Heyn, H. (2019). Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome biology* **20,** 110.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society:*

Series B (Methodological) **49(2),** 127–145.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions.* Marcel Dekker: New York.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Molecular cell* **58(4),** 610–620.

Li, C., Lu, J., Park, J., Kim, K., Brinkley, P. A., and Peterson, J. P. (1999). Multivariate zero-inflated poisson models and their applications. *Technometrics* **41(1),** 29–38.

Li, W. V. and Li, J. J. (2018). An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications* **9,** 997.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15(12),** 550.

Maher, M. J. (1990). A bivariate negative binomial model to explain traffic accident migration. *Accident Analysis & Prevention* **22(5),** 487–498.

Mc Mahon, S. S., Sim, A., Filippi, S., Johnson, R., Liepe, J., Smith, D., and Stumpf, M. P. (2014). Information theory and signal transduction systems: from molecular information processing to network inference. volume 35 of *Seminars in cell & developmental biology*, pages 98–108. Academic Press.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33(3),** 341–365.

Peng, T., Zhu, Q., Yin, P., and Tan, K. (2019). Scrabble: single-cell rna-seq imputation constrained by bulk rna-seq data. *Genome biology* **20,** 88.

Pont, F., Tosolini, M., and J, F. J. (2019). Single-cell signature explorer for comprehensive visualization of single cell signatures across scrna-seq data sets. *Nucleic Acids Research* **47,** e133.

Preisser, J. S., Das, K., Long, D. L., and Divaris, K. (2016). Marginalized zeroinflated negative binomial regression with application to dental caries. *Statistics in M edicine* **35(10),** 1722–1735.

Preisser, J. S., Long, D. L., and Stamm, J. W. (2017). Matching the statistical model to the research question for dental caries indices with many zero counts. *Caries research* **51,** 198–208.

R Core Team (2019). *R: A language and environment for statistical computing.*

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J. P. (2018). A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications* **9(1),** 284.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26(1),** 139–140.

van den Berge, K., Perraudeau, F., Soneson, C., Love, M. I., Risso, D., Vert, J. P., ..., and Clement, L. (2018). Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. *Genome biology* **19(1),** 24.

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* **174,** 716–729.

Wang, J., Huang, M., Torre, E., Dueck, H., Shaffer, S., Murray, J., Raj, A., Li, M., and Zhang, N. R. (2018). Gene expression distribution deconvolution in single-cell rna sequencing. *Proceedings of the National Academy of Sciences* **115,** E6437–E6446.

Wang, P. (2003). A bivariate zero-inflated negative binomial regression model for count data with excess zeros. *Economics Letters* **78(3),** 373–378.

Yu, T. (2018). A new dynamic correlation algorithm reveals novel functional aspects in single cell and bulk rna-seq data. *PLoS computational biology* **14,** e1006391.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4(1),**.

Supplementary Materials

Web Appendices A and B, referenced in Section 3, Web Appendix C, referenced in Section 7, Web Figures 1 to 5, referenced in Section 4 are available in the Supplementary Materials.

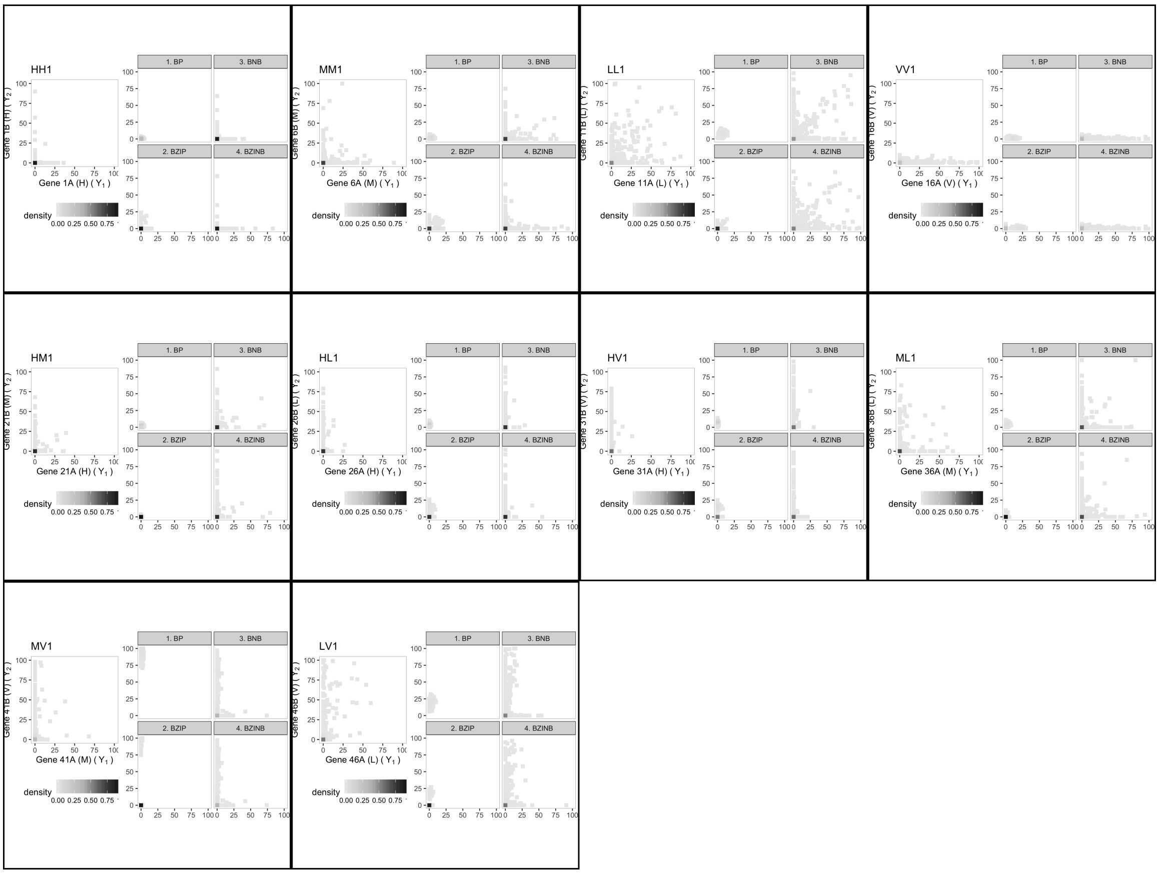
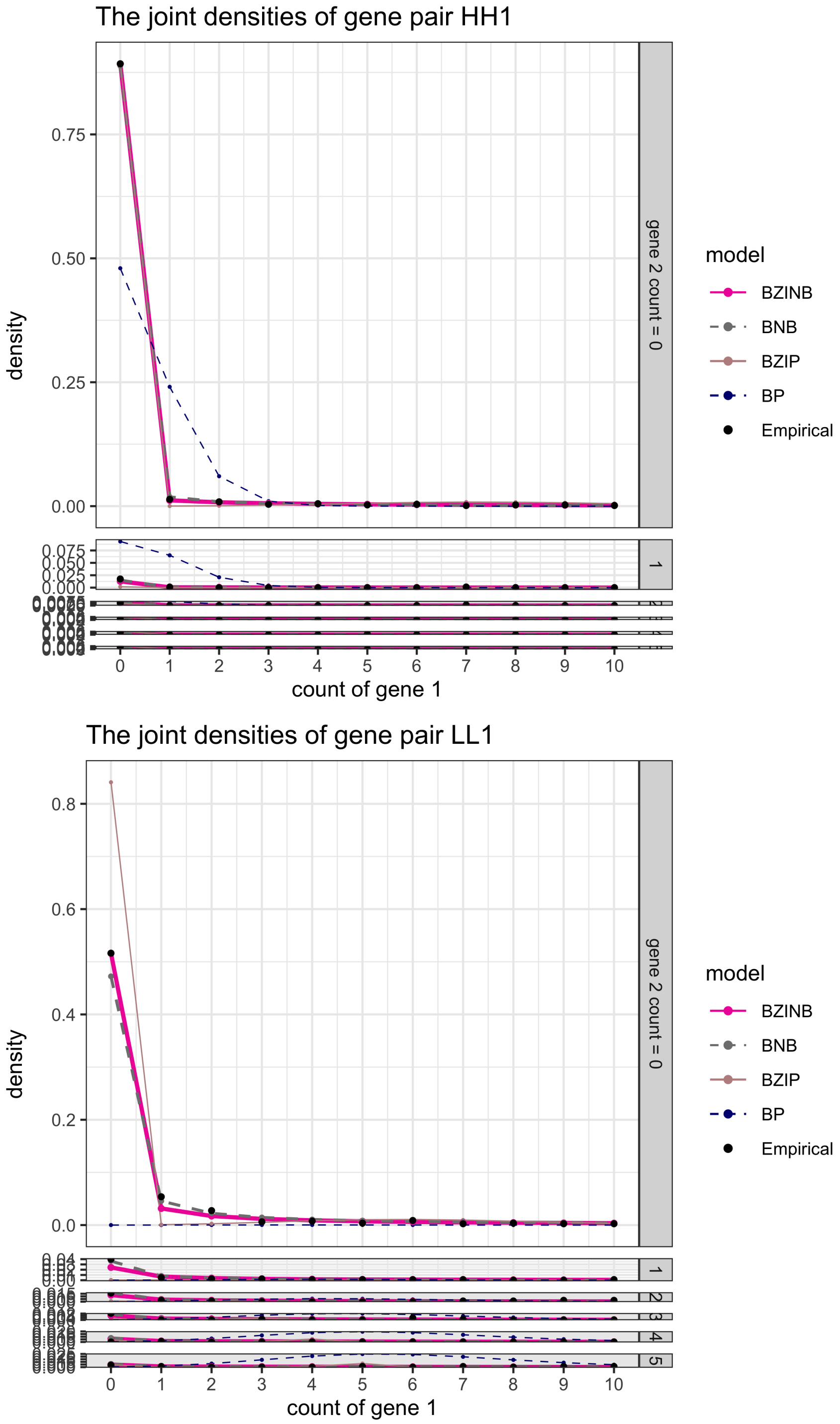


Figure 1. The bivariate distribution of real and simulated mouse paneth RNA count data. Each box corresponds to the first pair of each of the combination, HH1, MM1, LL1, VV1, HM1, HL1, HV1, ML1, MV1, and LV1, where letters represent stratum with varying proportions of zeros and the numbers represent the number of the pair in each combination. Each box has the real empirical distribution (LEFT), and the four model-based simulated empirical distributions (RIGHT).



The joint densities of gene pair HH1

The joint densities of gene pair LL1

Figure 2. The model estimates of bivariate densities (lines) and the empirical densities (dots) of two gene pairs.

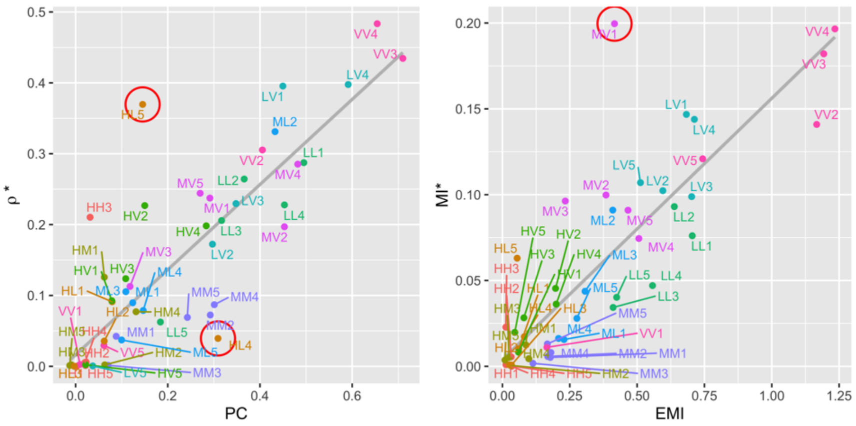


Figure 3. Estimated dependence measures of 50 pairs. Pearson correlation and underlying correlation estimates (LEFT). Empirical and underlying mutual information estimates (RIGHT).

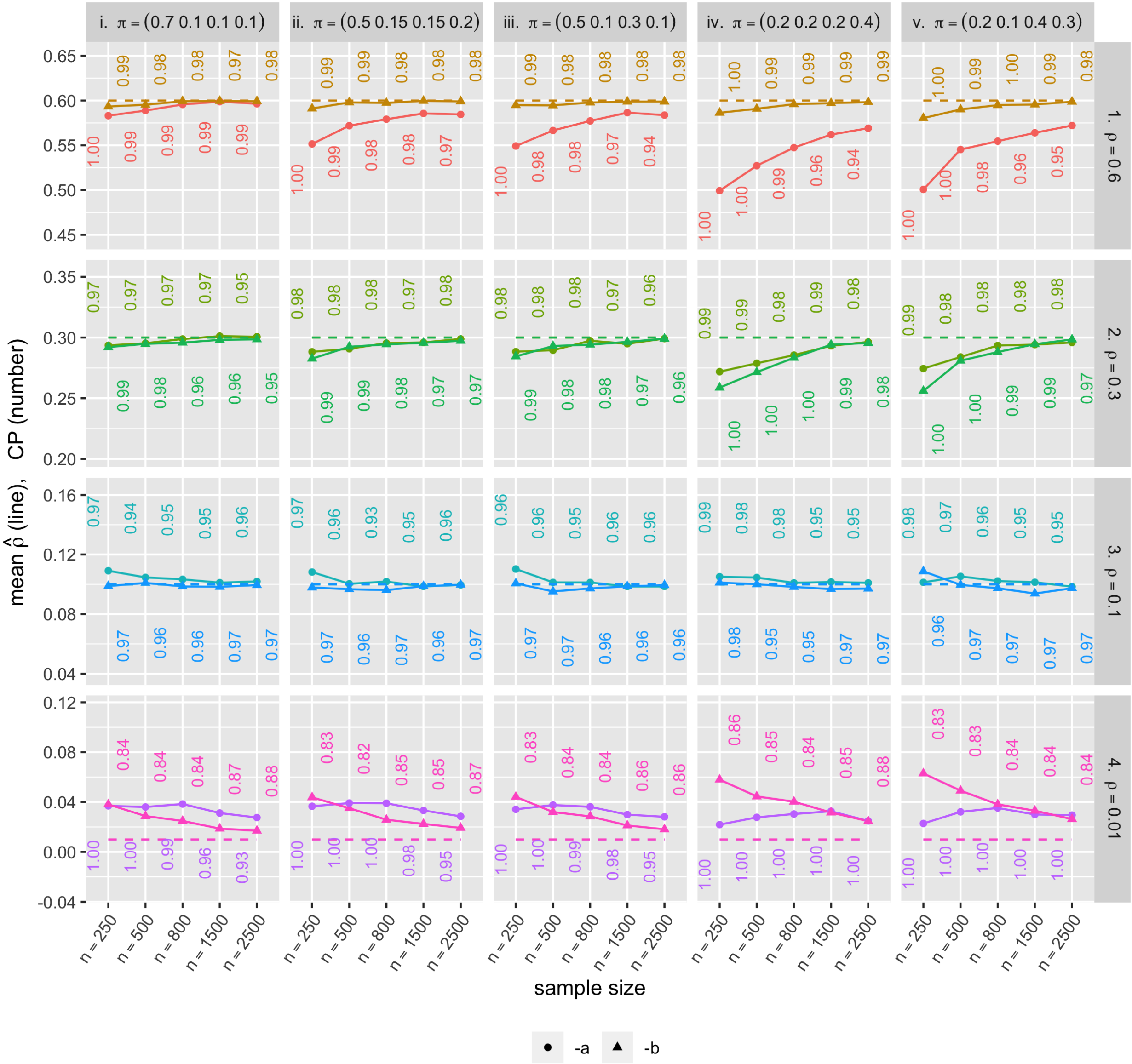


Figure 4. Mean parameter estimates ($\hat{\rho}^*$) and CP (each color represents distinct simulation scenarios.)

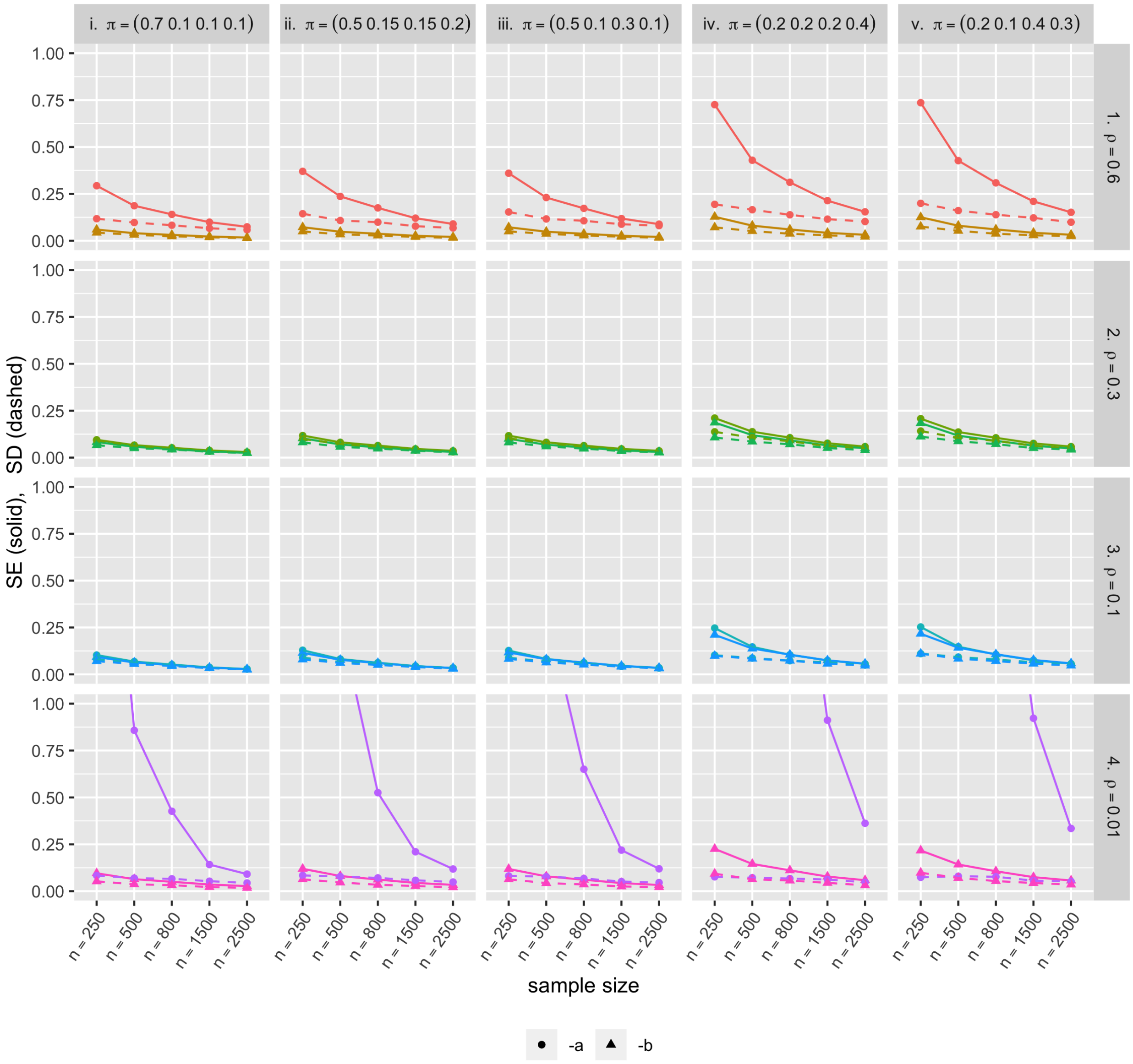


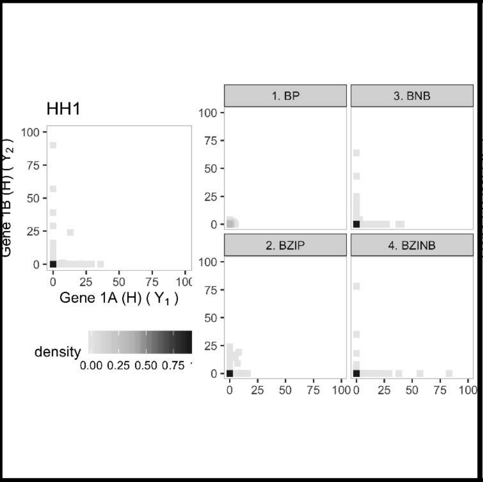
Figure 5. SE and SD (each color represents distinct simulation scenarios.)

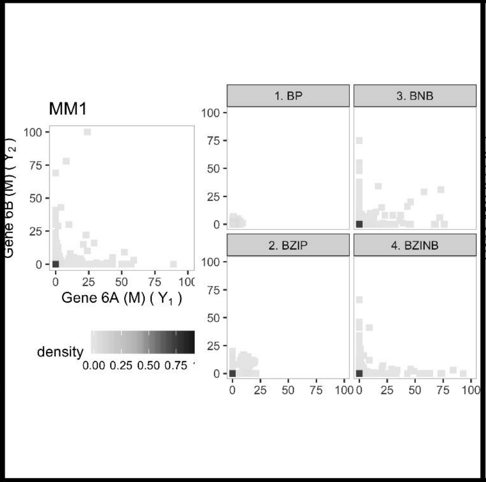
Table 1

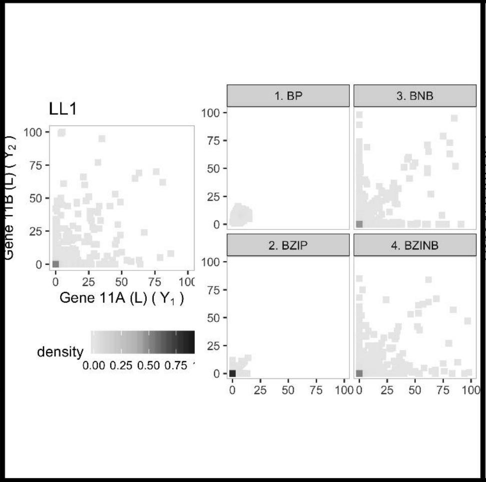
The set of parameters for simulation. Combination of $(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2)$ *and* $(\pi_1, \pi_2, \pi_3, \pi_4)$ *below makes* $40(= 8 \times 5)$ *sets in total.*

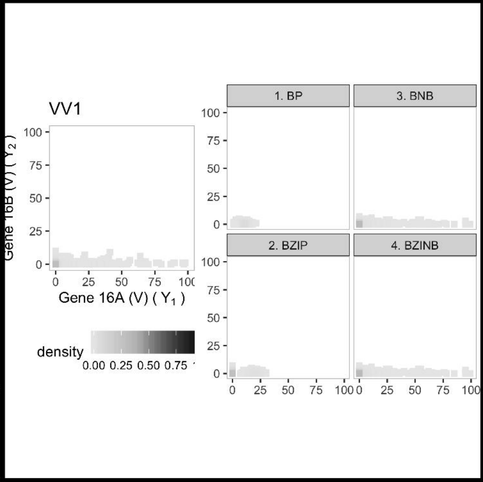
underlying correlation		#	$(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2)$
1. high	$(\rho^* = 0.6)$	1-a	(0.2, 0.05, 0.05, 3.0, 3.0)
		1-b	(2.0, 0.7, 0.1, 2.5, 2.5)
2. moderate	$(\rho^* = 0.3)$	2-a	(1.0, 1.0, 1.0, 1.5, 1.5)
		2-b	(3.0, 2.0, 1.0, 1.5, 0.5)
3. low	$(\rho^* = 0.1)$	3-a	(0.2, 0.3, 3.0, 2.0, 1.5)
		3-b	(0.5, 2.0, 2.0, 0.5, 3.0)
4. very low	$(\rho^* = 0.01)$	4-a	(0.01, 0.1, 1.0, 0.5, 0.5)
		4-b	(0.05, 2.0, 3.0, 3.0, 0.5)

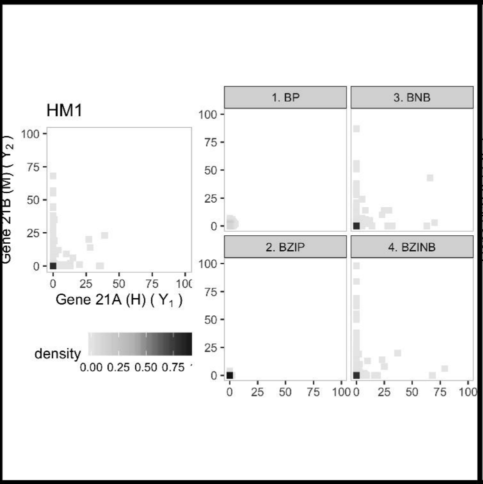
zero-inflation	$(\pi_1, \pi_2, \pi_3, \pi_4)$
i. low	(0.7, 0.1, 0.1, 0.1)
ii. moderate-balanced	(0.5, 0.15, 0.15, 0.2)
III. moderate-unbalanced	(0.5, 0.1, 0.3, 0.1)
iv. high-balanced	(0.2, 0.2, 0.2, 0.4)
v. high-unbalanced	(0.2, 0.1, 0.4, 0.3)

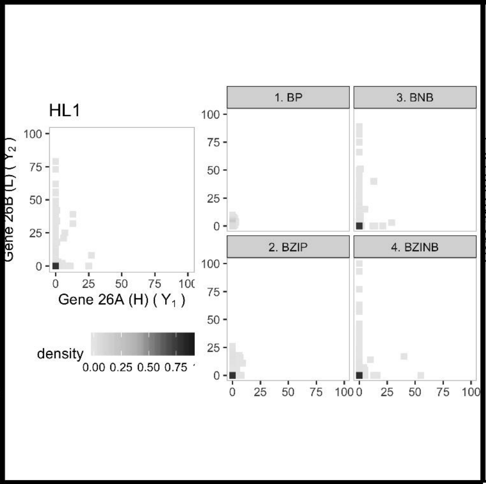

HH1
Gene 1B (H) (Y_2)
Gene 1A (H) (Y_1)
density
0.00 0.25 0.50 0.75
1. BP
3. BNB
2. BZIP
4. BZINB
0 25 50 75 100

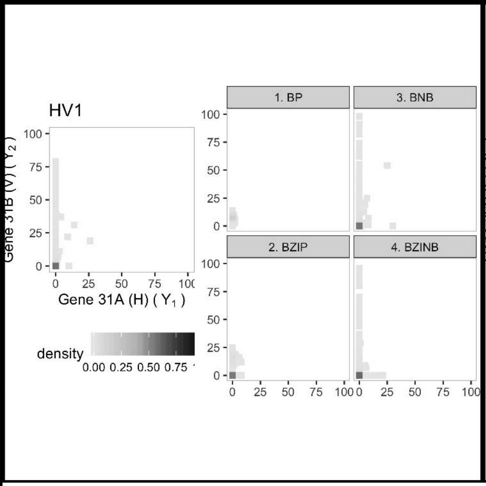

MM1
Gene 6B (M) (Y_2)
Gene 6A (M) (Y_1)
density
0.00 0.25 0.50 0.75
1. BP
3. BNB
2. BZIP
4. BZINB
0 25 50 75 100

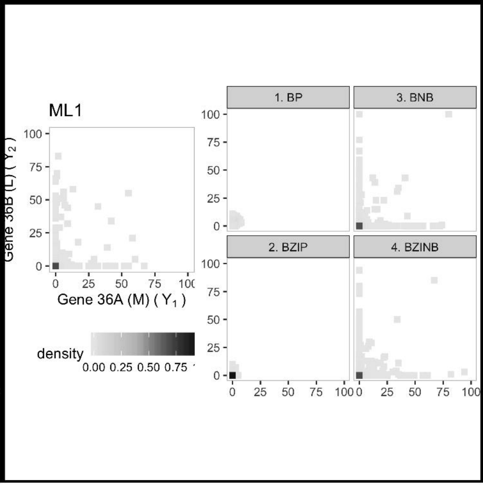

LL1
Gene 11B (L) (Y_2)
Gene 11A (L) (Y_1)
density
0.00 0.25 0.50 0.75
1. BP
3. BNB
2. BZIP
4. BZINB
0 25 50 75 100

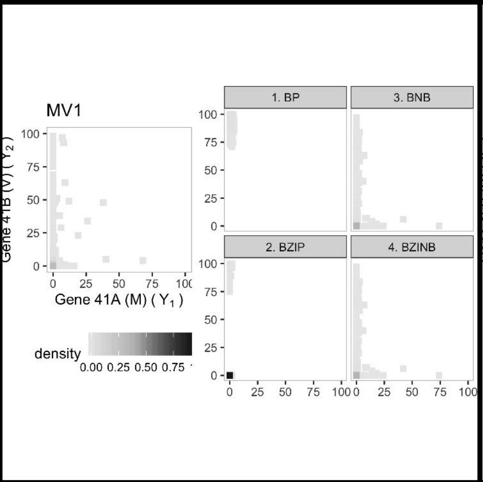

VV1
Gene 16B (V) (Y_2)
Gene 16A (V) (Y_1)
density
0.00 0.25 0.50 0.75
1. BP
3. BNB
2. BZIP
4. BZINB
0 25 50 75 100

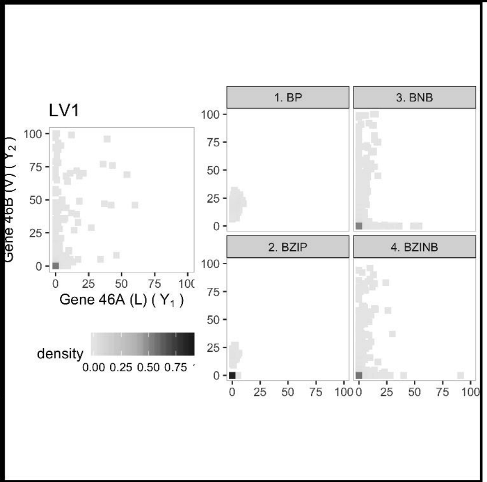

HM1
Gene 21B (M) (Y_2)
Gene 21A (H) (Y_1)
density
0.00 0.25 0.50 0.75
1. BP
3. BNB
2. BZIP
4. BZINB
0 25 50 75 100

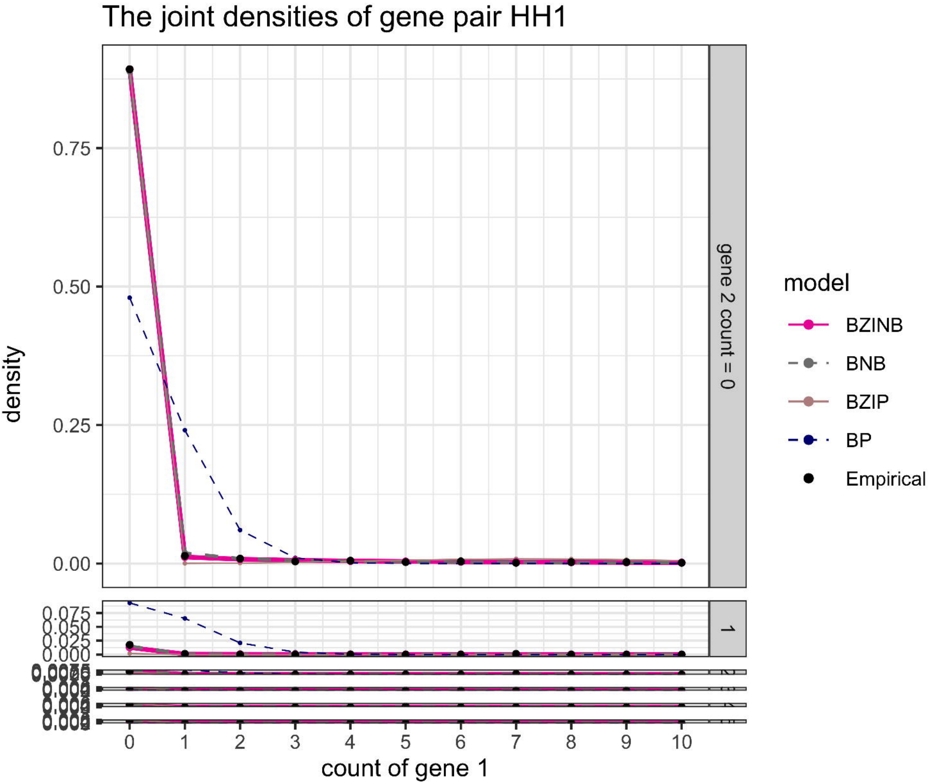

HL1
Gene 26B (L) (Y_2)
Gene 26A (H) (Y_1)
density
0.00 0.25 0.50 0.75
1. BP
3. BNB
2. BZIP
4. BZINB
0 25 50 75 100


HV1
Gene 31B (V) (Y_2)
Gene 31A (H) (Y_1)
density
0.00 0.25 0.50 0.75
1. BP
3. BNB
2. BZIP
4. BZINB
0 25 50 75 100

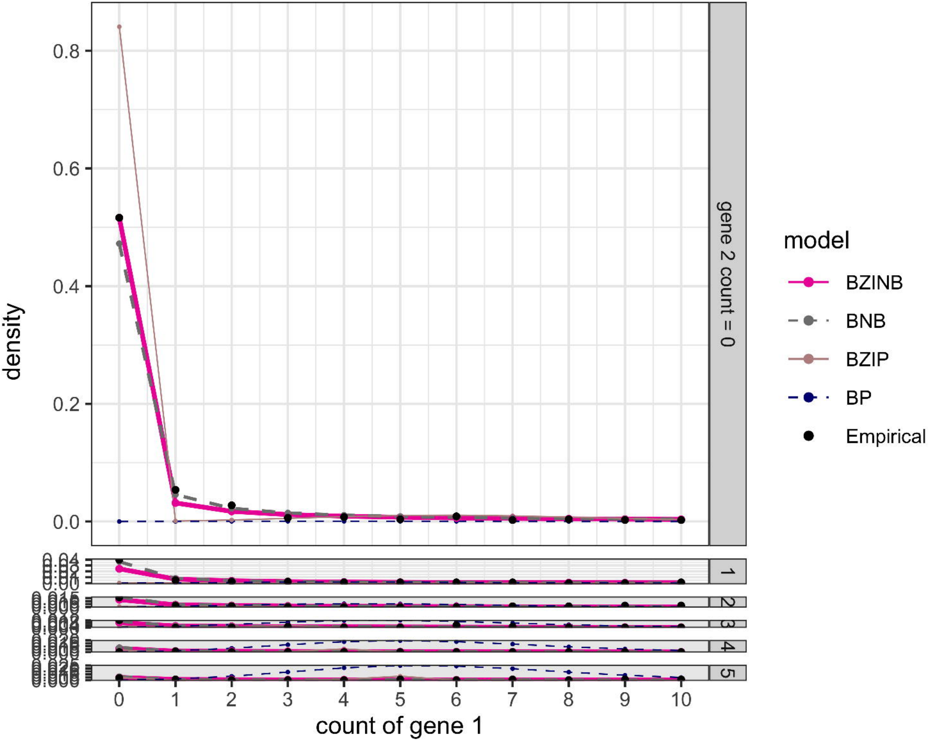

ML1
Gene 36B (L) (Y_2)
Gene 36A (M) (Y_1)
density
0.00 0.25 0.50 0.75
1. BP
3. BNB
2. BZIP
4. BZINB
0 25 50 75 100

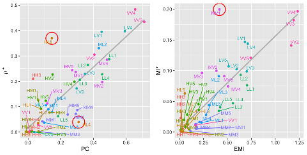

MV1
Gene 41B (V) (Y_2)
Gene 41A (M) (Y_1)
density
0.00 0.25 0.50 0.75
1. BP
3. BNB
2. BZIP
4. BZINB
0 25 50 75 100


LV1
Gene 46B (V) (Y_2)
Gene 46A (L) (Y_1)
density
0.00 0.25 0.50 0.75
1. BP
3. BNB
2. BZIP
4. BZINB
0 25 50 75 100

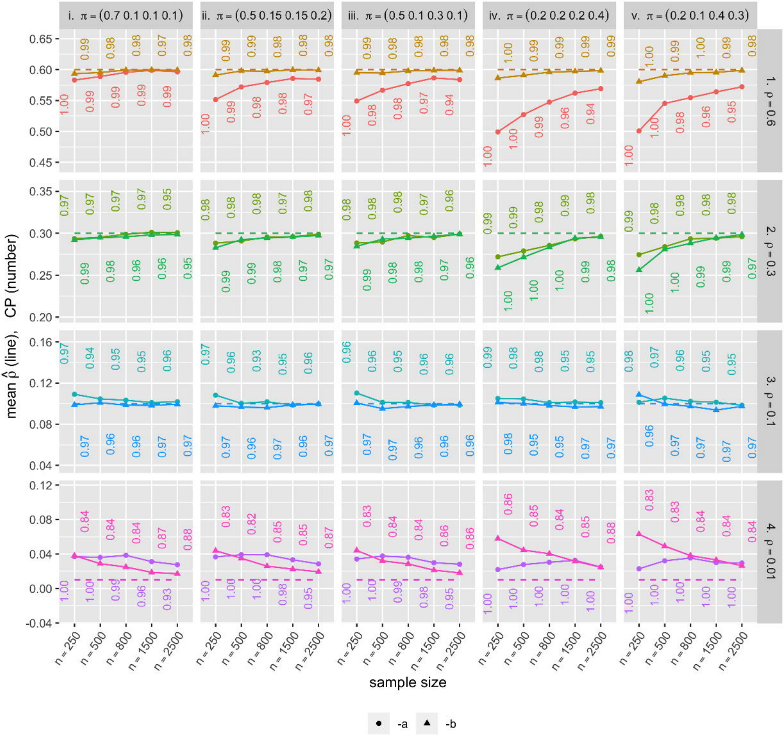

The joint densities of gene pair HH1
density
0.75
0.50
0.25
0.00
0.075
0.050
0.025
0.000
gene 2 count = 0
1
0
1
2
3
4
5
6
7
8
9
10
count of gene 1
model
BZINB
BNB
BZIP
BP
Empirical

The joint densities of gene pair LL1

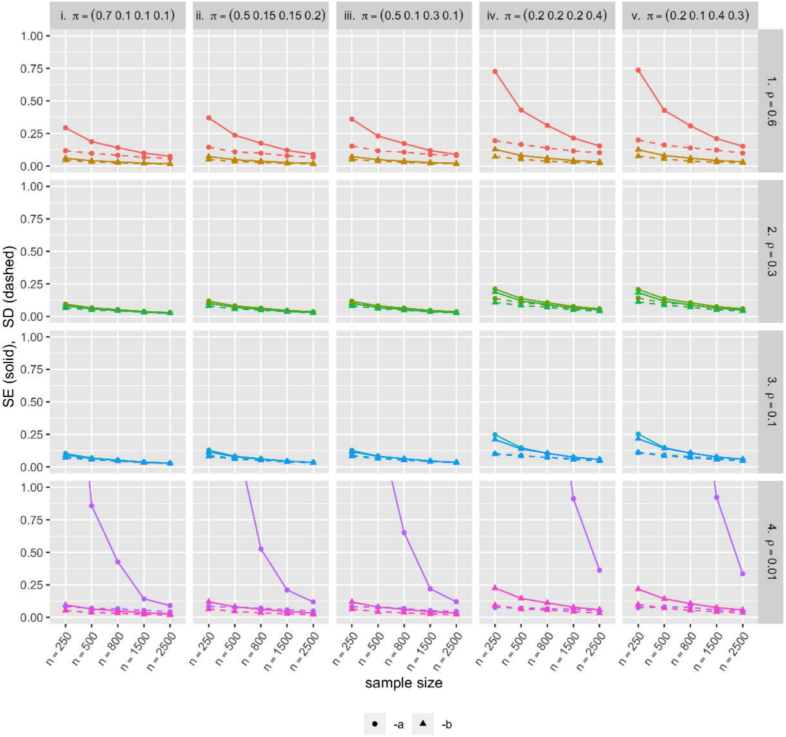




μ*
PC
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
VV4
VV3
LV4
LV1
HL5
ML2
LL1
VV2
MV4
MV5
LL2
MV1
LV3
LL4
HH3
HV2
HV4
LL3
MV2
LV2
HM1
MV3
HV1
HV3
ML4
HL1
ML3
MM1
MM5
MM4
VV1
HM4
LL5
ML4
HM2
MM3
HV5
HH5
HH3
MI*
EMI
0.00
0.05
0.10
0.15
0.20
0.00
0.25
0.50
0.75
1.00
1.25
MV1
VV4
VV3
LV1
LV4
VV2
VV5
LV5
LV2
LV3
MV2
MV3
ML2
MV5
LL2
HV5
HV2
LL1
HL5
HV3
HV4
MV4
ML3
HH3
HV1
LL5
LL4
HH2
ML5
LL3
HM3
MM5
HM1
ML1
VV1
MM2
MM1
MM4
HM2
MM3
HH1



i. π = (0.7 0.1 0.1 0.1)
ii. π = (0.5 0.15 0.15 0.2)
iii. π = (0.5 0.1 0.3 0.1)
iv. π = (0.2 0.2 0.2 0.4)
v. π = (0.2 0.1 0.4 0.3)
1. ρ = 0.6
2. ρ = 0.3
3. ρ = 0.1
4. ρ = 0.01
mean ρ̂ (line), CP (number)
0.65
0.60
0.55
0.50
0.45
0.35
0.30
0.25
0.20
0.16
0.12
0.08
0.04
0.12
0.08
0.04
0.00
-0.04
0.99 0.98 0.98 0.97 0.98
1.00 0.99 0.99 0.99 0.99
0.99 0.99 0.98 0.98 0.98
1.00 0.99 0.98 0.98 0.97
0.99 0.98 0.98 0.98 0.98
1.00 0.98 0.98 0.97 0.94
1.00 0.99 0.99 0.99 0.99
1.00 1.00 0.99 0.96 0.94
1.00 0.99 1.00 0.99 0.98
1.00 1.00 0.98 0.96 0.95
0.97 0.97 0.97 0.97 0.95
0.99 0.98 0.96 0.96 0.95
0.98 0.98 0.98 0.97 0.98
0.99 0.99 0.98 0.97 0.97
0.98 0.98 0.98 0.97 0.96
0.99 0.98 0.98 0.97 0.96
0.99 0.99 0.98 0.99 0.98
1.00 1.00 1.00 0.99 0.98
0.99 0.98 0.98 0.98 0.98
1.00 1.00 0.99 0.99 0.97
0.97 0.94 0.95 0.95 0.96
0.97 0.96 0.96 0.97 0.97
0.97 0.96 0.93 0.95 0.96
0.97 0.96 0.97 0.96 0.97
0.96 0.96 0.95 0.96 0.96
0.97 0.97 0.96 0.96 0.96
0.99 0.98 0.98 0.95 0.95
0.98 0.95 0.95 0.97 0.97
0.98 0.97 0.96 0.95 0.95
0.96 0.97 0.97 0.97 0.97
0.84 0.84 0.84 0.87 0.88
1.00 1.00 0.99 0.96 0.93
0.83 0.82 0.85 0.85 0.87
1.00 1.00 1.00 0.98 0.95
0.83 0.84 0.84 0.86 0.86
1.00 1.00 0.99 0.98 0.95
0.86 0.85 0.84 0.85 0.88
1.00 1.00 1.00 1.00 1.00
0.83 0.83 0.84 0.84 0.84
1.00 1.00 1.00 1.00 1.00
n = 250
n = 500
n = 800
n = 1500
n = 2500
sample size
● -a
▲ -b


i. π = (0.7 0.1 0.1 0.1)
ii. π = (0.5 0.15 0.15 0.2)
iii. π = (0.5 0.1 0.3 0.1)
iv. π = (0.2 0.2 0.2 0.4)
v. π = (0.2 0.1 0.4 0.3)
1. ρ = 0.6
2. ρ = 0.3
3. ρ = 0.1
4. ρ = 0.01
SE (solid), SD (dashed)
1.00
0.75
0.50
0.25
0.00
n = 250
n = 500
n = 800
n = 1500
n = 2500
sample size
● -a
▲ -b