Received: 28 August 2019 | Revised: 3 March 2021 | Accepted: 8 March 2021

DOI: 10.1111/biom.13457

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC PRACTICE

Modeling dynamic correlation in zero-inflated bivariate count data with applications to single-cell RNA sequencing data

Zhen Yang | **Yen-Yi Ho**

Department of Statistics, University of South Carolina, Columbia, South Carolina, USA

Correspondence
Zhen Yang, Department of Statistics, University of South Carolina, Columbia, SC, USA.
Email: zheny@email.sc.edu

Funding information
National Cancer Institute, Grant/Award Number: 1R21CA264353-01

Abstract

Interactions between biological molecules in a cell are tightly coordinated and often highly dynamic. As a result of these varying signaling activities, changes in gene coexpression patterns could often be observed. The advancements in next-generation sequencing technologies bring new statistical challenges for studying these dynamic changes of gene coexpression. In recent years, methods have been developed to examine genomic information from individual cells. Single-cell RNA sequencing (scRNA-seq) data are count-based, and often exhibit characteristics such as overdispersion and zero inflation. To explore the dynamic dependence structure in scRNA-seq data and other zero-inflated count data, new approaches are needed. In this paper, we consider overdispersion and zero inflation in count outcomes and propose a ZEro-inflated negative binomial dynamic COrrelation model (ZENCO). The observed count data are modeled as a mixture of two components: success amplifications and dropout events in ZENCO. A latent variable is incorporated into ZENCO to model the covariate-dependent correlation structure. We conduct simulation studies to evaluate the performance of our proposed method and to compare it with existing approaches. We also illustrate the implementation of our proposed approach using scRNA-seq data from a study of minimal residual disease in melanoma.

KEYWORDS

correlated count data, covariate-dependent correlation, dynamic coexpression, liquid association, single-cell RNA sequencing, zero inflation

1 | INTRODUCTION

Interactions between biological molecules in a cell are tightly coordinated and often highly dynamic (Luscombe *et al.*, 2004; de Lichtenberg *et al.*, 2005). They can change flexibly under different cellular conditions or in response to various external stimulants and signals. As a result of these varying signaling activities, changes in gene coexpression patterns can often be observed in these situations (Li, 2002; Li and Yuan, 2004; de la Fuente, 2010). Studying these dynamic changes in gene coexpression could reveal these intricate underlying gene regulatory mechanisms.

Although it is a challenging task to unravel the complex genetic interactions in a biological system, several

statistical approaches have been introduced to describe the coexpression between a pair of genes such as Pearson correlation or rank correlation, *F*-statistic (Lai *et al.*, 2004), mutual information (Faith *et al.*, 2007), entropy-based approaches (Ho *et al.*, 2007), Gaussian graphical models (Ma *et al.*, 2007), and Bayesian network (Ho *et al.*, 2014). However, these approaches do not account for the fact that genetic circuits can be turned on or off and genes may participate in different regulatory processes under different cellular conditions.

One statistical measure that can capture these dynamic gene correlation changes was proposed by Li (2002). This measure, named dynamic correlation in this paper, quantifies the relationship where the coexpression between two genes is modulated by a third "coordinator" gene. Li (2002) examined these dynamic correlation changes (referred to as liquid association in his paper) in canonical pathways using microarray gene expression data from a model organism, *Saccharomyces cerevisiae*. For a typical genomic study, a pathway-based or a genome-wide screening strategy can be implemented as presented in several studies to effectively identify potential dynamic correlation changes (Dawson and Kendziorski, 2012; Gunderson and Ho, 2014; Wang *et al.*, 2017; Yu, 2018; Kinzy *et al.*, 2019). Li's study and other studies since then have evidently established its biological validity and popularized it to be a useful tool for analyzing genomic data (Li, 2002; Li *et al.*, 2004; Ho *et al.*, 2007; Zhang *et al.*, 2007; Ho *et al.*, 2011; Wang *et al.*, 2013; Khayer *et al.*, 2017; Wang *et al.*, 2017; Xu *et al.*, 2017; Ai *et al.*, 2019; Kong and Yu, 2019; Wen *et al.*, 2020).

However, when it comes to count data such as RNA sequencing reads, these existing Gaussian-based approaches may not fit the data properly. RNA sequencing (RNA-seq) data are often presented as a count matrix with nonnegative counts as the number of reads observed. Count-based models such as the Poisson distribution and the negative binomial distribution are widely used to analyze the RNA-seq data. Karlis and Meligkotsidou (2005) proposed a multivariate Poisson model with covariance structure. Due to both biological and technical variability, RNA-seq count data are often overdispersed. For overdispersed data, the variance is larger than the mean, which is a violation of the assumption of the Poisson distribution (mean and variance are equal). To handle overdispersion, Solis-Trapala and Farewell (2005) used a multivariate Poisson-Gamma mixture model. Robinson *et al.* (2010) modeled the data using the negative binomial distribution and treated the Poisson distribution as a special case of the negative binomial distribution. Ma *et al.* (2020) proposed flexible models for modeling bivariate correlated count data.

In recent years, the rapid development of next-generation sequencing technologies has made it possible to examine the sequence information from individual cells. Single-cell RNA sequencing (scRNA-seq) analyzes the expression of RNAs from individual cells, whereas traditional RNA-seq can only analyze the RNAs from mixed cell populations (Bacher and Kendziorski, 2016; Hwang *et al.*, 2018). scRNA-seq gives insight into individual cells' function and behavior at various stages and in various cell types, and hence, can provide a high-resolution view of dynamic coexpression regulation in a biological system.

However, the analysis of scRNA-seq data is complicated by high levels of technical noise and intrinsic biological variability (Kharchenko *et al.*, 2014). Due to the low amounts of mRNA within individual cells, the counts of single-cell gene expression data contain a large number of zero expression measurements. To avoid stochastic zero counts, Lun *et al.* (2016) developed a normalization method based on pooling expression values. Pierson and Yau (2015) developed a dimensionality-reduction method considering the dropout characteristics to improves modeling accuracy. Miao *et al.* (2018) used a zero-inflated negative binomial model to estimate the proportion of real and dropout zeros. Kharchenko *et al.* (2014) modeled the measurement of each cell as a mixture of two components: one for transcripts that are successfully detected and the other for dropout events during amplification.

Motivated by the dynamic correlation studies in microarray data, in this article, we propose the ZEro-inflated negative binomial dynamic COrrelation (ZENCO) model. We account for overdispersion and zero inflation in count data by considering a mixture model of conditional bivariate negative binomial regressions and zero counts. A latent variable is incorporated into ZENCO to model the covariate-dependent correlation structure. We demonstrate the implementation of ZENCO model using the scRNA-seq data of melanoma cells from Gene Expression Omnibus (GSE116237) and study the difference of dynamic correlations between various phases during combined BRAF and MEK (BRAF/MEK) treatment.

The remainder of the article is arranged as follows. In Section 2, the detail of the proposed model is introduced. The simulation studies and comparisons are conducted in Section 3. In Section 4, the analysis of scRNA-seq data generated from melanoma tumor cells is presented. Section 5 concludes this article with some discussion.

2 | METHOD

2.1 | The ZENCO model

For modeling dynamic coexpression changes, we use $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ to denote the count-based expression levels for three genes. Let X_{ij} represent the gene expression level of the ith gene ($i = 1, 2, 3$) in the jth cells ($j = 1, 2, \ldots n$), and $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, \ldots, X_{in})$ represents

the gene expression level for the ith gene. In our proposed framework, the marginal distribution of $\mathbf{X}_i$ is modeled as a mixture of dropout component and negative binomial component (nondropout events). The distribution of $\mathbf{X}_i$ is given by

$$\mathbf{X}_i \sim \begin{cases} \mathcal{I}_0, & \text{with probability } p_i; \\ NB(\mu_i, \phi_i), & \text{with probability } 1 - p_i. \end{cases} \quad (1)$$

where $\mathcal{I}_0$ is the distribution with a point mass at zero; p_i is the dropout rate of $\mathbf{X}_i$; μ_i is the mean of the negative binomial component of $\mathbf{X}_i$; and ϕ_i is the dispersion parameter of the negative binomial component. The variance of the negative binomial component of $\mathbf{X}_i$ is $\mu_i(1 + \phi_i\mu_i)$. As ϕ_i goes to 0, $NB(\mu_i, \phi_i) \to Poisson(\mu_i)$.

The dropout rate of a given gene, p_i, is modeled as a function of its mean. The dropout rates are study-specific and can be estimated for a given scRNA-seq data set. Based on the melanoma data considered in the study, we model the dropout rate using a logistic function: $p = \frac{e^{(b_0+b_1\mu)}}{1+e^{(b_0+b_1\mu)}}$, where μ is the mean of a given gene and b_0, b_1 can be estimated using the expression levels of all available genes in the data (Pierson and Yau, 2015).

Furthermore, we use the indicator $d_{ij} \sim Bernoulli(p_i)$ to describe whether dropout happens or not. If $d_{ij} = 0$, then the ith gene in the jth cell is successfully amplified (nondropout event). If $d_{ij} = 1$, then dropout happens. According to the combinations of different values of d_{1j} and d_{2j}, there are four different situations for $\mathbf{X}_1$ and $\mathbf{X}_2$. Their marginal densities can be written as:

$$\begin{cases} X_{1j} \sim NB(\mu_1, \phi_1) \text{ and } X_{2j} \sim NB(\mu_2, \phi_2), & \text{if } d_{1j} = d_{2j} = 0; \\ X_{1j} \sim \mathcal{I}_0 \text{ and } X_{2j} \sim NB(\mu_2, \phi_2), & \text{if } d_{1j} = 1 \text{ and } d_{2j} = 0; \\ X_{1j} \sim NB(\mu_1, \phi_1) \text{ and } X_{2j} \sim \mathcal{I}_0, & \text{if } d_{1j} = 0 \text{ and } d_{2j} = 1; \\ X_{1j} \sim \mathcal{I}_0 \text{ and } X_{2j} \sim \mathcal{I}_0, & \text{if } d_{1j} = d_{2j} = 1. \end{cases} \quad (2)$$

When $d_{1j} = d_{2j} = d_{3j} = 0$, the joint distribution of $\mathbf{X}_1$ and $\mathbf{X}_2$ involves a correlation parameter that depends on the expression level of X_{3j}. In other words, the correlation between X_{1j} and X_{2j} could change according to the level of X_{3j} when all three genes (X_{1j}, X_{2j}, and X_{3j}) are successfully amplified in the jth cell. If d_{1j}=1 or $d_{2j} = 1$, X_{1j} and X_{2j} are independent, because at least one measurement of X_{1j} and X_{2j} comes from the dropout component.

We model the dependency between $\mathbf{X}_1$ and $\mathbf{X}_2$ and construct our conditional bivariate negative binomial model through a Poisson–Gamma mixture distribution. For $i = 1, 2$ and $j = 1, 2, \ldots, n$, let

$$X_{ij} \sim Poisson(u_{ij}\mu_i), u_{ij} \sim Gamma(\alpha_i, \alpha_i). \quad (3)$$

A negative binomial distribution of $NB(\mu_i, \frac{1}{\alpha_i})$ can be generated by integrating over u_{ij} in (3). In this Poisson–Gamma mixture setting, u_{ij} can be considered as the cell-specific random effect. To introduce the conditional correlation between X_{1j} and X_{2j} given X_{3j}, we utilize a latent variable Z and model the conditional correlation implicitly through the cell-specific random effect (u_{ij}).

Let $\mathbf{Z_j} = (Z_{1j}, Z_{2j})'$ be a bivariate normal variable that

$$\mathbf{Z_j} \sim N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho_j \\ \rho_j & 1\end{bmatrix}\right). \quad (4)$$

The correlation, ρ_j, of (Z_{1j}, Z_{2j}) is specified as

$$\log\left(\frac{1+\rho_j}{1-\rho_j}\right) = \tau_0 + \tau_1 X_{3j}. \quad (5)$$

$\log(\frac{1+\rho_j}{1-\rho_j})$ is the Fisher's Z-transformation for the correlation ρ_j that ensures that the correlation ρ_j is within $(-1, 1)$.

Now, we incorporate this latent variable $\mathbf{Z_j}$ into the cell-specific random component (u_{ij}) in the Poisson–Gamma mixture in (3) to construct a conditional bivariate negative binomial model of $(X_{1j}, X_{2j})'$ with marginal distribution $X_{1j} \sim NB(\mu_1, \phi_1)$ and $X_{2j} \sim NB(\mu_2, \phi_2)$ and the correlation of (X_{1j}, X_{2j}) depends on X_{3j}. Specifically, for $i = 1, 2$ and $j = 1, 2, \ldots, n$, let

$$X_{ij} \sim Poisson[F_{\alpha_i}^{-1}\{\Phi(Z_{ij})\}\mu_i], \quad (6)$$

where $F_{\alpha_i}(\cdot)$ is the cumulative distribution function of a $Gamma(\alpha_i, \alpha_i)$ distribution with $\alpha_i = 1/\phi_i$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. $F_{\alpha_i}^{-1}$ maps each point in the interval (0,1) to $Gamma(\alpha_i, \alpha_i)$ distribution. Hence, the distribution of $F_{\alpha_i}^{-1}\{\Phi(Z_{ij})\}$ is $Gamma(\alpha_i, \alpha_i)$. The distribution of $X_{ij} \sim Poisson[F_{\alpha_i}^{-1}\{\Phi(Z_{ij})\}\mu_i]$ is then a Poisson–Gamma mixture distribution, which follows the negative binomial density $NB(\mu_i, \phi_i = \frac{1}{\alpha_i})$.

In the model described above, in order to determine the existence of the dynamic coexpression change of $\mathbf{X}_1$, $\mathbf{X}_2$ given $\mathbf{X}_3$, the main parameter of interest is τ_1 in (5). If τ_1=0, then the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ does not depend on $\mathbf{X}_3$ and vice versa. In the ZENCO model, we develop a statistical inference procedure via a Bayesian perspective, because it offers a relatively straightforward way to compute $Poisson[F_{\alpha_i}^{-1}\{\Phi(Z_{ij})\}]$ through Markov chain Monte Carlo (MCMC) sampling. In addition, the posterior distributions of the parameters can be obtained with a set of standard conjugate priors.

Under the hypotheses:

$$H_0 : \tau_1 = 0 \text{ versus } H_1 : \tau_1 \neq 0,$$

the statistical power of the proposed ZENCO approach can be calculated as follows. First, we obtained the posterior sampling distribution of τ_1, and then calculated the 95% equal tail credible interval. Power can be evaluated as the proportion of times when zero is not covered by the 95% credible intervals.

We now describe the likelihood function and the MCMC scheme. Let vector $\boldsymbol{\theta}$ be the notation of all parameters (μ_1, μ_2, μ_3, ϕ_1, ϕ_2, ϕ_3, τ_0, τ_1) in the model. And let $\boldsymbol{\pi}(\boldsymbol{\theta})$ be the prior joint distribution of $\boldsymbol{\theta}$, the likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\theta}|x_1, x_2, x_3) &= \prod_{j=1}^{n} f(x_{1j}, x_{2j}|\mu_1, \mu_2, \phi_1, \phi_2, \tau_0, \tau_1, x_{3j}) f(x_{3j}|\mu_3, \phi_3) \\ &= \prod_{j=1}^{n} \left\{ \int f(x_{1j}, x_{2j}|\mu_1, \mu_2, \phi_1, \phi_2, \boldsymbol{z}_j) f(\boldsymbol{z}_j|x_{3j}, \tau_0, \tau_1) d\boldsymbol{z}_j \right\} f(x_{3j}|\mu_3, \phi_3) \\ &= \prod_{j=1}^{n} \left\{ \int \prod_{i=1}^{2} f(x_{ij}|\mu_i, \phi_i, \boldsymbol{z}_{ij}) f(\boldsymbol{z}_j|x_{3j}, \tau_0, \tau_1) d\boldsymbol{z}_j \right\} f(x_{3j}|\mu_3, \phi_3), \end{aligned} \tag{7}$$

where x_{1j} and x_{2j} are from observed data and $\boldsymbol{z}_j = (z_{1j}, z_{2j})'$. x_{1j} and x_{2j} are independent given $\boldsymbol{z}_j$. Hence, the posterior joint distribution of μ_1, μ_2, μ_3, ϕ_1, ϕ_2, ϕ_3, τ_0, τ_1 given the observations is proportional to

$$\left[\prod_{j=1}^{n} \left\{ \int \prod_{i=1}^{2} f(x_{ij}|\mu_i, \phi_i, \boldsymbol{z}_{ij}) f(\boldsymbol{z}_j|x_{3j}, \tau_0, \tau_1) d\boldsymbol{z}_j \right\} f(x_{3j}|\mu_3, \phi_3) \right] \boldsymbol{\pi}(\boldsymbol{\theta}),$$

where $f(x_{ij}|\mu_i, \phi_i, \boldsymbol{z}_{ij})$ is the distribution of x_{ij} for $i = 1, 2$:

$$x_{ij} \sim \begin{cases} \mathcal{I}_0, & \text{with probability } p_i; \\ \text{Poisson}\left[F^{-1}_{1/\phi_i}\{\Phi(z_{ij})\}\mu_i\right], & \text{with probability } 1 - p_i. \end{cases}$$

The dropout rate p_i is study-specific and can be determined using all genes measured in the study as a function of μ_i described previously. And $f(\boldsymbol{z}_j|x_{3j}, \tau_0, \tau_1)$ is the probability density function of a bivariate normal distribution with a covariance matrix structure:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \frac{e^{(\tau_0+\tau_1\times x_{3j})} - 1}{e^{(\tau_0+\tau_1\times x_{3j})} + 1} \\ \frac{e^{(\tau_0+\tau_1\times x_{3j})} - 1}{e^{(\tau_0+\tau_1\times x_{3j})} + 1} & 1 \end{bmatrix}.$$

For any given x_{3j}, $\boldsymbol{z}_j$ can be derived as described in (4) and (5). Finally, $f(x_{3j}|\mu_3, \phi_3)$ is formulated as in (1).

For a given gene triplet, the parameter estimation can be carried out using the MCMC algorithm provided in JAGS (Plummer, 2003). We use the normal distribution with mean 0 and variance $4/N$ as the priors of τ_0 and τ_1, where N is the sample size. This is because the approximate variance of Fisher's Z-transformation $\log(\frac{1+\rho}{1-\rho})$ is $\frac{4}{N-3}$. The priors for μ_1, μ_2, and μ_3 are standard log-normal distributions. The noninformative priors for the dispersion parameters $1/\phi_1$, $1/\phi_2$, and $1/\phi_3$ are the Gamma distribution with mean 100 and relatively large variance 10,000.

The sampling scheme during each MCMC iteration is as follows. For $j = 1, 2, \ldots, n$, $i = 1, 2, 3$, we sample μ_i from $f(\mu_i|\cdot) \propto f(\mu_i) \prod_{j=1}^{n} f(x_{ij}|\mu_i, \phi_i)$ and sample ϕ_i from $f(1/\phi_i|\cdot) \propto f(1/\phi_i) \prod_{j=1}^{n} f(x_{ij}|\mu_i, \phi_i)$, where $f(x_{ij}|\mu_i, \phi_i)$ is the probability density function of

$$x_{ij} \sim \begin{cases} \mathcal{I}_0, & \text{with probability } p_i; \\ NB(\mu_i, \phi_i), & \text{with probability } 1 - p_i. \end{cases}$$

Then we sample τ_0 from

$$f(\tau_0|\cdot) \propto f(\tau_0) \prod_{j=1}^{n} f(\mathbf{z_j}|\tau_0, \tau_1, x_{3j}),$$

and sample τ_1 from

$$f(\tau_1|\cdot) \propto f(\tau_1) \prod_{j=1}^{n} f(\mathbf{z_j}|\tau_0, \tau_1, x_{3j}),$$

where

$$f(\mathbf{z_j}|\tau_0, \tau_1, x_{3j}) = N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \frac{e^{(\tau_0+\tau_1\times x_{3j})} - 1}{e^{(\tau_0+\tau_1\times x_{3j})} + 1} \\ \frac{e^{(\tau_0+\tau_1\times x_{3j})} - 1}{e^{(\tau_0+\tau_1\times x_{3j})} + 1} & 1 \end{bmatrix} \right).$$

In addition, z_{ij} can be sampled from

$$f(z_{ij}|\cdot) \propto f(x_{ij}|z_{ij}, \mu_i, \alpha_i) f(z_{ij}|z_{kj}), i, k = 1, 2; i \neq k,$$

where $f(z_{ij}|z_{kj}) = N(\rho_j z_{kj}, (1 - {\rho_j}^2))$.

2.2 | Search strategies

There are several ways to implement the ZENCO approach in a genomic study. We describe a few here: (i) for a given pair of genes ($\mathbf{X}_1$, $\mathbf{X}_2$), screen the whole genome to identify the coordinator genes ($\mathbf{X}_3$) that regulate the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$, or (ii) for a given $\mathbf{X}_3$, screen-related pathways or the whole genome to identify pairs of genes that are modulated by $\mathbf{X}_3$ (m choose 2 gene pairs; m is the total number of genes considered), or (iii) if no prior information about $\mathbf{X}_3$ or ($\mathbf{X}_1$, $\mathbf{X}_2$) is available, screen relevant genetic pathways, or screen the whole genome to identify potential gene triplets that exhibit dynamic correlation changes (m choose three gene triplets). In the experimental data analysis described in Section 4, we demonstrated the second (ii) approach.

When the number of relevant genes under consideration is large (for example, $\approx$ 20,000), a prescreening step is usually beneficial before implementing ZENCO. For example, the algorithm proposed by Gunderson and Ho (2014) or the screening statistic (ζ) introduced in Yu (2018) or filtering out gene with constant expression has been used effectively in the literature.

3 | SIMULATION

To evaluate the performance of our proposed ZENCO model and compare it to existing benchmark approaches, we report results from five simulation scenarios below.

3.1 | Scenario 1: Simulating data from ZENCO

In this first simulation, we demonstrate generating data from the ZENCO model. The simulated data contain count-based expression level of three genes: $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$. In our model, the correlations of $\mathbf{X}_1$ and $\mathbf{X}_2$ are modulated by the level of $\mathbf{X}_3$. This simulation was conducted as follows.

First, we simulated a set of $\{x_{3j}\}_{j=1}^N$ from a univariate negative binomial distribution with mean μ_3 and size ϕ_3 and then randomly selected a subset as the dropouts and replaced these $\{x_{3j}\}'s$ with zero. After the simulation of x_{3j}, we calculated correlation coefficient $\rho_j = \frac{e^{(\tau_0+\tau_1\times x_{3j})}-1}{e^{(\tau_0+\tau_1\times x_{3j})}+1}$ for each x_{3j}. Note that for dropouts in $\{x_{3j}\}_{j=1}^N$, we used μ_3 instead of x_{3j} to calculate ρ_j, because the values of those dropouts have nothing to do with the regulatory mechanism of $\mathbf{X}_3$. Then, we generated latent variables $\mathbf{z}_j = (z_{1j}, z_{2j})'$ such that

$$\mathbf{z}_j \sim N_2\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho_j\\ \rho_j & 1\end{bmatrix}\right)$$

and simulated x_{1j} and x_{2j} using $\mathbf{z}_j$ as described in (6). The dependence structure of x_{1j} and x_{2j} is implicitly modeled via $\mathbf{z}_j$. Finally, just like the simulation of x_{3j}, we randomly replaced values of x_{1j} and x_{2j} for dropout events.

Using the simulation approach described above, we generated 10^5 observations from the ZENCO distribution and plotted a panel of conditional distributions of $\mathbf{X}_1$ and $\mathbf{X}_2$ given various levels of $\mathbf{X}_3$ in Figure 1. In these figures, we observed that when $\mathbf{X}_3$ is not zero, ρ increases with $\mathbf{X}_3$. When $\mathbf{X}_3$ is zero, the correlations of $\mathbf{X}_1$ and $\mathbf{X}_2$ are small and show reduced dependency with respect to $\mathbf{X}_3$. This is due to the zero value observation of $\mathbf{X}_3$ being a mixture of true zero and dropout. In other words, some zero values of $\mathbf{X}_3$ come from the negative binomial distribution, others come from dropout events.

3.2 | Scenario 2: Comparisons to existing approaches

To evaluate the performance of our proposed ZENCO model, we performed power analysis and compare ZENCO to three other existing approaches. For testing the existence of dynamic coexpression changes, our hypotheses are set up as:

$$H_0 : \tau_1 = 0 \text{ versus } H_1 : \tau_1 \neq 0.$$

First, we compared ZENCO to a bivariate negative binomial regression without considering the zero-inflated components. Similarly to ZENCO, the statistical power of this method can be calculated as the percentage of times that the posterior 95% credible intervals of τ_1 do not cover zero. The ZENCO model and the model without considering the zero-inflated components were both carried out using the MCMC algorithm with 20,000 iterations, and 10,000 burn-ins.

Second, we compared ZENCO to the existing benchmark approach introduced by Li (2002). This existing approach was later applied to scRNA-seq data by Yu (2018). This test statistic according to the three-product-moment measure is written as: $T_{LA} = \frac{\hat{E}(\mathbf{X}_1^*\mathbf{X}_2^*\mathbf{X}_3^*)}{SE\{\hat{E}(\mathbf{X}_1^*\mathbf{X}_2^*\mathbf{X}_3^*)\}}$, where $\mathbf{X}_1^*$, $\mathbf{X}_2^*$, $\mathbf{X}_3^*$ are the standardized $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$ with mean 0, variance

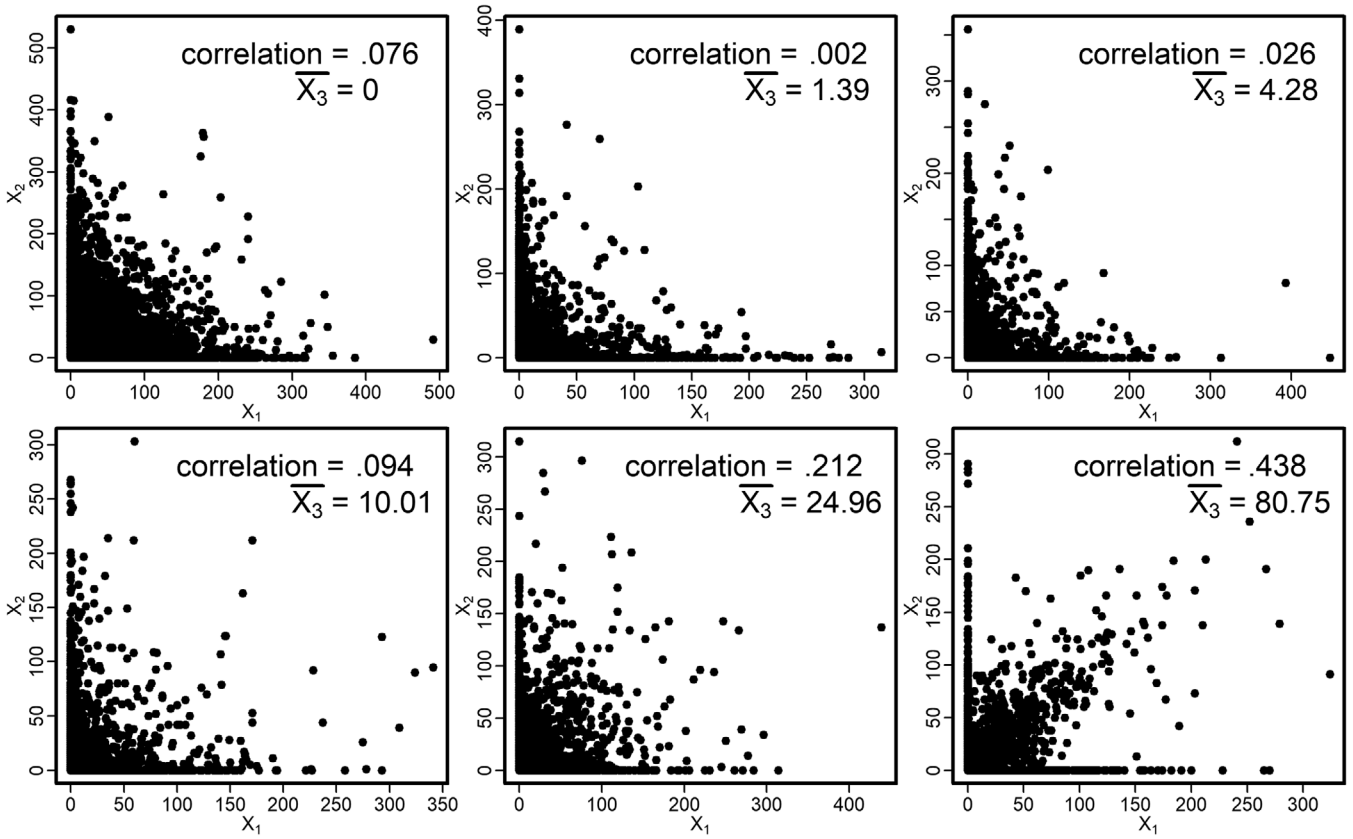

FIGURE 1 Profile plots of $(\mathbf{X}_1, \mathbf{X}_2|\mathbf{X}_3)$ with varying $\mathbf{X}_3$ ($\mu_1 = \mu_2 = \mu_3 = 15$, $\phi_1 = \phi_2 = \phi_3 = 4$, $\tau_0 = 0$, and $\tau_1 = 0.05$)

1, and $\widehat{E}(\mathbf{X}_1^*\mathbf{X}_2^*\mathbf{X}_3^*)$ is the three-product-moment estimator for the dynamic correlation. $SE\{\widehat{E}(\mathbf{X}_1^*\mathbf{X}_2^*\mathbf{X}_3^*)\}$, the standard error of $\widehat{E}(\mathbf{X}_1^*\mathbf{X}_2^*\mathbf{X}_3^*)$, can be estimated via bootstrap. T_{LA} can be used to test whether the correlation of $\mathbf{X}_1$, $\mathbf{X}_2$ depends on $\mathbf{X}_3$, that is, $H_0 : \tau_1 = 0$ (Li, 2002; Ho *et al.*, 2011). The distribution of T_{LA} under the null hypothesis and associated p-value can be obtained using a permutation approach.

The third comparison is to fit the negative binomial count data with the conditional normal model (CNM-Full) (Ho *et al.*, 2011). Assuming that data are from the conditional bivariate normal distribution instead of the conditional bivariate negative binomial distribution, the test statistic of this method can be estimated using a generalized estimating equation-based procedure (Yan and Fine, 2004) and a p-value associated with the test statistic can be obtained. The powers of these two methods (T_{LA} and CNM-Full) can be calculated by counting the percentage of times when p-values associated with τ_1 are less than .05.

We simulated 1000 observations from ZENCO model by fixing $\mu_1 = \mu_2 = \mu_3 = 15$, $\phi_1 = \phi_2 = \phi_3 = 4$, and $\tau_0 = 0$, and then varied τ_1 values and performed power analyses. The simulated values of $\mu_1, \mu_2, \mu_3, \phi_1, \phi_2, \phi_3$ are based on the estimates obtained from the real data analysis.

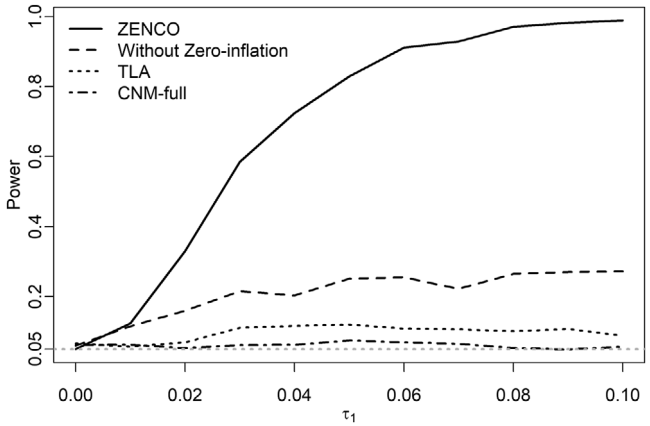

FIGURE 2 Power curves comparing various methods. Both TLA and CNM-Full approaches are Gaussian-based models

Figure 2 shows the power curves of the four methods. We observed that our proposed ZENCO method outperforms the other three methods. In addition, fitting the negative binomial count-based data using Gaussian-based models reduces statistical power drastically. This is because ZENCO accounts for both zero inflation and overdispersion of the data, and hence achieves better power to detect dynamic dependence structure.

TABLE 1 Coverage probability of 95% credible intervals (CIs) and interval lengths based on 1000 MCMC simulations ($\tau_0 = 0.01$, $\tau_1 = 0.05$)

		Without zero inflation		With zero inflation	
	Parameter	Coverage probability	CI length	Coverage probability	CI length
$N = 200$	τ_0	1.000	0.237	1.000	0.246
	τ_1	0.154	0.041	0.957	0.095
$N = 500$	τ_0	1.000	0.223	1.000	0.244
	τ_1	0.006	0.022	0.961	0.059
$N = 1000$	τ_0	0.957	0.205	1.000	0.242
	τ_1	0.000	0.015	0.954	0.040

TABLE 2 Mean square errors (MSEs) and mean bias errors (MBEs) based on 1000 MCMC simulations ($\tau_0 = 0.01$, $\tau_1 = 0.05$)

		Without zero inflation		With zero inflation	
	Parameter	MSE	MBE	MSE	MBE
$N = 200$	τ_0	0.001	0.005	0.000	−0.008
	τ_1	0.002	−0.039	0.001	−0.006
$N = 500$	τ_0	0.002	0.024	0.000	−0.009
	τ_1	0.002	−0.040	0.000	−0.001
$N = 1000$	τ_0	0.004	0.048	0.000	−0.009
	τ_1	0.002	−0.041	0.000	0.000

3.3 | Scenario 3: Estimation efficiency

In this simulation scenario, we evaluated the estimation efficiency of the ZENCO model and reported mean squared errors (MSE), mean bias errors (MBE), and 95% empirical coverage probabilities under various settings. Three sets of simulation studies were done with sample sizes 200, 500, and 1000. For each simulation study, we generated 1000 data sets. We used the parameter estimated values obtained from the real data analysis in Section 4 and set the true values of the parameters as follows: $\mu_1 = \mu_2 = \mu_3 = 15$, $\phi_1 = \phi_2 = \phi_3 = 4$, $\tau_0 = 0.01$, and $\tau_1 = 0.05$. The true values of the parameters associated with dropout rate were similar to the values obtained based on the real data: $b_0 = 0.14$ and $b_1 = -0.02$ (dropout rates for $\mathbf{X}_1$ and $\mathbf{X}_2$ are both 0.44).

The empirical 95% coverage probabilities from the posterior distributions and the length of credible intervals are shown in Table 1. In Table 1, we also presented the parameter estimates using a negative binomial model without zero inflation. The empirical 95% coverage probability is calculated as the percentage of times when the 95% credible intervals covering the true parameter value based on 1000 MCMC simulations. The simulation results shown in Table 1 suggest that ZENCO model provides a much better 95% coverage probability than a negative binomial regression method model without zero inflation.

MSEs and MBEs are shown in Table 2. The MBE of a given parameter β is calculated as $\frac{1}{N}\sum_{i=1}^{N}(\widehat{\beta}_i - \beta)$; N is the number of simulation iterations ($N = 1000$). Based on the simulation results in Table 2, ZENCO model has smaller MSEs and MBEs comparing with the nonzero-inflated negative binomial regression method.

3.4 | Scenario 4: Robustness

To assess the robustness of the ZENCO method under model misspecification, we conducted three sets of simulations where the data are generated via a negative binomial model without zero inflation. The three sets of simulation studies were performed with sample sizes 200, 500, and 1000, and each with 1000 simulation iterations. The true values of parameters were set as $\mu_1 = \mu_2 = \mu_3 = 15$, $\phi_1 = \phi_2 = \phi_3 = 4$, $\tau_0 = 0.01$, and $\tau_1 = 0.05$. We analyzed the simulated data sets using a negative binomial regression method without zero inflation and the ZENCO method.

The empirical 95% coverage probabilities from posterior distributions and the length of credible intervals using the above two models are shown in Table S.1; the MSEs and MBEs are shown in Table S.2. The simulation results shown in Table S.1 and Table S.2 suggest that our proposed estimation procedure in ZENCO is fairly robust even when the data are generated from a nonzero-inflated negative binomial setting.

3.5 | Scenario 5: A multiple-gene setting

In this simulation scenario, we turn our attention to a multiple-gene setting. Our goal here is to demonstrate that our proposed approach could capture dependencies among multiple genes through multiple pairwise searches. We set $b_0 = 0.65$ and $b_1 = -0.015$, which is similar to the values obtained based on the real data and then simulated five genes (10 gene pair combinations) with $\mu_1 = 15$, $\mu_2 = 19, \mu_3 = 10, \mu_4 = 15, \mu_5 = 12, \phi_1 = 4, \phi_2 = 5, \phi_3 = 6$, $\phi_4 = 4$, $\phi_5 = 3$. The true values of the 10 $\tau_1's$ range from 0.005 to 0.05, whereas the true value of τ_0 was set as 0. The empirical 95% coverage probabilities and MBEs of 10 $\tau_1's$ are shown in Table S.3. The results indicate that our method demonstrated desirable performance under a multiple-gene setting.

4 | EXPERIMENTAL DATA ANALYSIS

We used the proposed ZENCO model to analyze the melanoma data set described in Rambow *et al.* (2018). The scRNA-seq data were obtained from Gene Expression Omnibus (GEO accession number: GSE116237). The data set consists of 57,445 genes and 674 melanoma cells. To study minimal residual disease (MRD) as well as relapse during melanoma treatment, Rambow *et al.* (2018) performed scRNA-seq using malignant cells from BRAF-mutant patient-derived xenograft melanoma cohorts treated with BRAF/MEK inhibitor (dabrafenib/trametinib).

During the course of continuous treatment with BRAF/MEK inhibitor, the transition of tumor cells can be categorized into three phases: phase 1 is in the early stage when all treated lesions rapidly shrunk upon initial treatment (BRAF-inhibitor sensitive); phase 2 is the second stage when drug-tolerant tumor cells remain viable upon continuous treatment (MRD); in phase 3, relapse is observed and tumor cells exhibit adaptive resistance to continuous BRAF inhibition treatment (BRAF-inhibitor resistance). Among the 674 melanoma cells in the data set, there are 155 phase 1 cells, 199 phase 2 cells, and 148 phase 3 cells. More details can be found in Rambow *et al.* (2018).

To gain insight into transcriptional switches of genetic circuits in tumor cells during the course of BRAF-inhibitor treatment, we set out to identify gene pairs that interact with BRAF differently between BRAF-inhibitor sensitive cells (phase 1) and BRAF-inhibitor resistance cells (phase 3). Hence, in this analysis, we chose BRAF as $\mathbf{X}_3$ and conducted the pairwise analysis for genes in the melanoma pathway described in the KEGG database (Kanehisa and Goto, 2000). According to the melanoma pathway in KEGG database, 72 genes were identified as melanoma-associated genes. The data were first preprocessed by the procedures described in McCarthy *et al.* (2017). After removing low expressed genes (maximum count across all cells less than 5) and genes with more than 70% zeros in either phase 1 cells or phase 3 cells, 28 genes were selected for further analysis.

The study-specific parameters, b_0, b_1, associated with dropout rates can be estimated using the logistic function $p = \frac{e^{(b_0+b_1\mu)}}{1+e^{(b_0+b_1\mu)}}$. In the logistic function, we used the sample mean to estimate μ. After calculating the dropout rate as the proportion of cells with zero counts, a nonlinear least-squares approach was then applied to calculate b_0 and b_1.

We implemented ZENCO analyses for 351 gene pair combinations in phase 1 cells and phase 3 cells and obtained the estimates of τ_1. To identify the gene pairs that interact with BRAF differently, we chose gene pairs that are in both phase 1 and phase 3 cells and calculated the differences of τ_1 estimates between the two phases. The top 30 gene pairs with the largest differences of τ_1 between phase 3 and phase 1 are shown in Table 3.

The first two columns in Table 3 are the names of two genes. $\tau_1(P1)$ is the estimated τ_1 in phase 1 cells, and $\tau_1(P3)$ is the estimated τ_1 in phase 3 cells. $\Delta\tau_1$ is defined as $\tau_1(P3) - \tau_1(P1)$. It quantifies the change of dynamic coexpression in relation to BRAF between phase 3 and phase 1 cells.

From Table 3, we observed that genes PDGFC and FGFR1 have the largest $|\Delta\tau_1|$ between phase 1 and phase 3 cells. In phase 1 cells, the estimate of τ_1 for PDGFC and FGFR1 is 0.045 and the 95% credible interval does not contain 0. In phase 3 cells, the estimate of τ_1 is close to 0. This suggests that the regulatory mechanism between BRAF and the gene pair (PDGFC, FGFR1) changes between phase 1 and phase 3 cells. Czyz (2019) pointed out that melanoma cells somehow acquire the ability to grow independent of the two growth factors: FGFR1, PDGFC that helps melanoma cells to gain resistance toward BRAF treatment. Our finding from Table 3 is consistent with this finding. Interestingly, many top gene pairs listed in Table 3 are from the mitogen-activated protein kinase (MAPK) and phosphoinositide 3-kinase (PI3K) signaling pathways. Our analysis findings support the hypotheses described in Villanueva *et al.* (2011).

In the above analysis, the convergence of MCMC was assessed using the Gelman–Rubin convergence statistic (Gelman *et al.*, 1992). The convergence statistics were close to 1 for all τ_1 estimates in all 351 gene pairs. The trace plots of the top five gene pairs are shown in Figure S.1. In our real data application, it took 67 minutes to implement ZENCO with three chains (100,000 iterations each) for all 351 gene

TABLE 3 Top table of dynamic correlations differences. $\Delta\tau_1$ is the difference between τ_1 estimates in phase 3 (P3) and phase 1 (P1)

#	Gene1	Gene2	$\tau_1(P1)$	$\tau_1(P3)$	$\Delta\tau_1$
1	PDGFC	FGFR1	0.045 (0.021, 0.068)	−0.003 (−0.010, 0.005)	−0.047 (−0.072,−0.023)
2	AKT1	BAX	0.040 (0.008, 0.071)	−0.003 (−0.014, 0.008)	−0.043 (−0.075,−0.010)
3	AKT1	PIK3R1	−0.016 (−0.035, 0.004)	0.024 (0.009, 0.038)	0.040 (0.015, 0.062)
4	PDGFC	MAP2K2	0.016 (−0.002, 0.032)	−0.023 (−0.036,−0.006)	−0.039 (−0.059,−0.013)
5	IGF1R	FGFR1	−0.024 (−0.048, 0.000)	0.007 (0.000, 0.014)	0.032 (0.006, 0.056)
6	MDM2	CCND1	0.021 (0.007, 0.031)	−0.011 (−0.018,−0.004)	−0.031 (−0.044,−0.017)
7	AKT1	ARAF	−0.025 (−0.047, 0.002)	0.007 (−0.007, 0.018)	0.031 (0.002, 0.056)
8	AKT1	MAP2K1	0.025 (0.004, 0.057)	−0.006 (−0.017, 0.009)	−0.030 (−0.063,−0.006)
9	AKT1	MAPK1	−0.003 (−0.012, 0.006)	0.026 (0.007, 0.055)	0.029 (0.007, 0.058)
10	KRAS	PDGFC	0.012 (−0.005, 0.024)	−0.017 (−0.042, 0.005)	−0.029 (−0.057,−0.002)
11	IGF1R	MAP2K2	0.025 (0.002, 0.056)	−0.004 (−0.011, 0.006)	−0.028 (−0.060,−0.004)
12	PTEN	PDGFC	−0.022 (−0.036,−0.004)	0.007 (−0.003, 0.014)	0.028 (0.008, 0.044)
13	PTEN	PIK3R1	0.031 (0.007, 0.050)	0.005 (−0.006, 0.014)	−0.027 (−0.048,−0.002)
14	BAX	POLK	0.025 (0.006, 0.048)	0.000 (−0.012, 0.010)	−0.026 (−0.051,−0.003)
15	KRAS	NRAS	0.017 (−0.003, 0.034)	−0.008 (−0.015, 0.002)	−0.024 (−0.043,−0.003)
16	ARAF	RB1	0.020 (0.008, 0.032)	−0.004 (−0.009, 0.002)	−0.024 (−0.037,−0.011)
17	AKT1	RAF1	−0.016 (−0.033,−0.003)	0.007 (−0.004, 0.017)	0.023 (0.006, 0.042)
18	NRAS	MAPK1	0.017 (0.002, 0.029)	−0.005 (−0.013, 0.006)	−0.021 (−0.037,−0.004)
19	PIK3R1	MDM2	0.020 (0.004, 0.035)	−0.001 (−0.010, 0.008)	−0.021 (−0.038,−0.002)
20	IGF1R	TP53	−0.016 (−0.034, 0.002)	0.005 (−0.003, 0.011)	0.020 (0.002, 0.039)
21	BAK1	POLK	−0.018 (−0.030,−0.006)	0.002 (−0.006, 0.010)	0.020 (0.006, 0.034)
22	AKT3	MAP2K2	0.016 (0.005, 0.025)	−0.003 (−0.011, 0.007)	−0.018 (−0.030,−0.006)
23	PTEN	KRAS	−0.005 (−0.016, 0.011)	0.012 (0.003, 0.020)	0.017 (0.000, 0.030)
24	BAD	RAF1	−0.016 (−0.031,−0.006)	0.000 (−0.009, 0.008)	0.016 (0.002, 0.032)
25	IGF1R	CDK6	0.014 (−0.001, 0.026)	−0.002 (−0.008, 0.003)	−0.016 (−0.029,−0.001)
26	RB1	CCND1	0.011 (0.000, 0.020)	−0.004 (−0.010, 0.004)	−0.014 (−0.025,−0.002)
27	AKT2	FGFR1	−0.003 (−0.015, 0.006)	0.011 (0.004, 0.017)	0.014 (0.002, 0.027)
28	BAD	TP53	−0.001 (−0.010, 0.007)	0.013 (0.002, 0.021)	0.014 (0.001, 0.026)
29	NRAS	BAK1	0.001 (−0.008, 0.008)	0.014 (0.006, 0.022)	0.014 (0.002, 0.025)
30	AKT2	BAK1	−0.004 (−0.013, 0.005)	0.010 (0.000, 0.019)	0.014 (0.001, 0.026)

combinations using 13 computing cluster nodes (each with 28 2.4 GHz Intel Xeon E5-2680 v4 processors).

5 | DISCUSSION

In this paper, we presented a zero-inflated negative binomial dynamic correlation model for studying covariate-dependent correlations in zero-inflated, overdispersed count data, such as scRNA-seq data. In our model, the correlation of two genes is regulated by the expression level of the third gene; a phenomenon we named dynamic correlation in this paper. This novel dynamic correlation focuses on studying the changes of conditional correlation. It is a different measure from the partial correlation coefficient. The partial correlation quantifies the amount of residual correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ after regression on $\mathbf{X}_3$ to adjust for the influence of $\mathbf{X}_3$ (Li, 2002).

The proposed model in this paper takes both overdispersion and zero inflation of the data into consideration. With the proper choice of the values of parameters τ_0 and τ_1, the relationship between conditional correlation and the expression level of the third gene can be positive or negative. As demonstrated by our simulation studies, the ZENCO model significantly outperforms other existing approaches.

Two other prior distributions for the dispersion parameters ϕ_1, ϕ_2, and ϕ_3 have been implemented: an informative Gamma distribution on $\frac{1}{\phi}$ and a half-t-distribution on $\sqrt{\phi}$. Our sensitivity analysis suggests that the ϕ_1, ϕ_2, and ϕ_3 estimates are robust regardless of prior distribution assumptions. The Gamma distribution with mean 100 and

relatively large variance 10,000 used in this paper is more general and has slightly better performance in MCMC parameter estimates.

Moreover, in our model, ρ is the correlation of the latent variable Z. The Fisher transformation of ρ is assumed to be linear with X_3. In a more general setting, the relationship between $\log(\frac{1+\rho}{1-\rho})$ and $\mathbf{X}_3$ does not have to be linear. And our model can be easily adapted to other settings.

In the melanoma data analysis, $\mathbf{X}_3$ was used to denote the expression level of BRAF. And ZENCO model was implemented for each pairwise combination of $\mathbf{X}_1$ and $\mathbf{X}_2$ in the KEGG melanoma pathway. Using this search strategy, we found the pairs of genes whose BRAF-associated dynamic correlations change significantly between different phases during treatment. In Table 3, we reported the top genes with the largest $|\Delta\tau_1|$. Several existing type I error control approaches can be used in conjunction with the Bayesian model framework in ZENCO such as Käll *et al.* (2008) and Dawson and Kendziorski (2012). As described in Section 2, there are several ways to implement ZENCO in a genomic study. If a prefiltering step is used before implementing ZENCO, considerations described in van Iterson *et al.* (2010); Dawson and Kendziorski (2012) could be helpful to maintain type I error control.

Furthermore, in our application, $\mathbf{X}_3$ was used to denote the gene expression level of the BRAF gene because of its pivotal role in melanoma treatment and relapse in the study. In practice, the $\mathbf{X}_3$ can be easily modified to represent the activity level of a biological process or different cell types, or various cellular conditions such as tumor status, survival probability, degree of inflammation, metastasis potential, and so on. Also, $\mathbf{X}_3$ can be easily extended to represent a linear combination of several covariates or biological processes to accommodate the complexity of biological systems in other applications.

Because several existing procedures are available for preprocessing scRNA-seq data to remove low-magnitude background noise, in the ZENCO model, the dropout component is modeled as a degenerate distribution with a point mass at zero. However, the method can be easily adapted to allow a low-magnitude Poisson distribution to model the background noise in the dropout component.

In this paper, our focus is on the changes in coexpression patterns between a gene pair. It is plausible that there might exist higher order interactions between genes (more than two genes), and a generalization of our approach to higher dimensions is feasible. However, special treatments need to be considered to guarantee the positive definiteness of the variance–covariance matrix in higher dimension.

ORCID

Zhen Yang https://orcid.org/0000-0002-9344-9629
Yen-Yi Ho https://orcid.org/0000-0002-3224-3184

REFERENCES

Ai, D., Li, X., Pan, H., Chen, J., Cram, J.A. and Xia, L.C. (2019) Explore mediated co-varying dynamics in microbial community using integrated local similarity and liquid association analysis. *BMC Genomics*, 20, 185.

Bacher, R. and Kendziorski, C. (2016) Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biology*, 17, 63.

Czyz, M. (2019) Fibroblast growth factor receptor signaling in skin cancers. *Cells*, 8, 540.

Dawson, J.A. and Kendziorski, C. (2012) An empirical Bayesian approach for identifying differential coexpression in high-throughput experiments. *Biometrics*, 68, 455–465.

de la Fuente, A. (2010) From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends in Genetics : TIG*, 26, 326–333.

de Lichtenberg, U., Jensen, L.J., Brunak, S. and Bork, P. (2005) Dynamic complex formation during the yeast cell cycle. *Science*, 307, 724–727.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G. et al. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5, e8.

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–472.

Gunderson, T. and Ho, Y.-Y. (2014) An efficient algorithm to explore liquid association on a genome-wide scale. *BMC Bioinformatics*, 15, 371.

Ho, Y.-Y., Cope, L., Dettling, M. and Parmigiani, G. (2007) Statistical methods for identifying differentially expressed gene combinations. *Methods in Molecular Biology*, 408, 171–191.

Ho, Y.-Y., Cope, L.M. and Parmigiani, G. (2014) Modular network construction using eQTL data: an analysis of computational costs and benefits. *Frontiers in Genetics*, 5, 40.

Ho, Y.-Y., Parmigiani, G., Louis, T.A. and Cope, L.M. (2011) Modeling liquid association. *Biometrics*, 67, 133–141.

Hwang, B., Lee, J.H. and Bang, D. (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, 50, 1–14.

Käll, L., Storey, J.D., MacCoss, M.J. and Noble, W.S. (2008) Posterior error probabilities and false discovery rates: two sides of the same coin. *Journal of Proteome Research*, 7, 40–44.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28, 27–30.

Karlis, D. and Meligkotsidou, L. (2005) Multivariate poisson regression with covariance structure. *Statistics and Computing*, 15, 255–265.

Kharchenko, P.V., Silberstein, L. and Scadden, D.T. (2014) Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11, 740–742.

Khayer, N., Marashi, S.-A., Mirzaie, M. and Goshadrou, F. (2017) Three-way interaction model to trace the mechanisms involved in Alzheimer's disease transgenic mice. *PLoS One*, 12, e0184697.

Kinzy, T.G., Starr, T.K., Tseng, G.C. and Ho, Y.-Y. (2019) Meta-analytic framework for modeling genetic coexpression dynamics. *Statistical Applications in Genetics and Molecular Biology*, 18, 1–12.

Kong, Y. and Yu, T. (2019) A hypergraph-based method for large-scale dynamic correlation study at the transcriptomic scale. *BMC Genomics*, 20, 397.

Lai, Y., Wu, B., Chen, L. and Zhao, H. (2004) A statistical method for identifying differential gene–gene co-expression patterns. *Bioinformatics*, 20, 3146–3155.

Li, K.-C. (2002) Genome-wide coexpression dynamics: theory and application. *Proceedings of the National Academy of Sciences*, 99, 16875–16880.

Li, K.-C., Liu, C.-T., Sun, W., Yuan, S. and Yu, T. (2004) A system for enhancing genome-wide coexpression dynamics study. *Proceedings of the National Academy of Sciences*, 101, 15561–15566.

Li, K.-C. and Yuan, S. (2004) A functional genomic study on NCI's anticancer drug screen. *The Pharmacogenomics Journal*, 4, 127–135.

Lun, A.T., Bach, K. and Marioni, J.C. (2016) Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17, 75.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A. and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, 431, 308–312.

Ma, S., Gong, Q. and Bohnert, H.J. (2007) An *Arabidopsis* gene network based on the graphical Gaussian model. *Genome Research*, 17, 1614–1625.

Ma, Z., Hanson, T.E. and Ho, Y.-Y. (2020) Flexible bivariate correlated count data regression. *Statistics in Medicine*, 39, 3476–3490.

McCarthy, D.J., Campbell, K.R., Lun, A.T. and Wills, Q.F. (2017) Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, 33, 1179–1186.

Miao, Z., Deng, K., Wang, X. and Zhang, X. (2018) DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, 34, 3223–3224.

Pierson, E. and Yau, C. (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, 16, 1–10.

Plummer, M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Rambow, F., Rogiers, A., Marin-Bejar, O., Aibar, S., Femel, J., Dewaele, M. et al. (2018) Toward minimal residual disease-directed therapy in melanoma. *Cell*, 174, 843–855.

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, 139–140.

Solis-Trapala, I.L. and Farewell, V.T. (2005) Regression analysis of overdispersed correlated count data with subject specific covariates. *Statistics in Medicine*, 24, 2557–2575.

van Iterson, M., Boer, J.M. and Menezes, R.X. (2010) Filtering, FDR and power. *BMC Bioinformatics*, 11, 450.

Villanueva, J., Vultur, A. and Herlyn, M. (2011) Resistance to BRAF inhibitors: unraveling mechanisms and future treatment options. *Cancer Research*, 71, 7137–7140.

Wang, L., Liu, S., Ding, Y., Yuan, S., Ho, Y.-Y. and Tseng, G.C. (2017) Meta-analytic framework for liquid association. *Bioinformatics*, 33, 2140–2147.

Wang, L., Zheng, W., Zhao, H. and Deng, M. (2013) Statistical analysis reveals co-expression patterns of many pairs of genes in yeast are jointly regulated by interacting loci. *PLoS Genetics*, 9, e1003414.

Wen, X., Gao, L. and Hu, Y. (2020) LAcemodule: identification of competing endogenous RNA modules by integrating dynamic correlation. *Frontiers in Genetics*, 11, 235.

Xu, X., Wang, M., Li, L., Che, R., Li, P., Pei, L. and Li, H. (2017) Genome-wide trait-trait dynamics correlation study dissects the gene regulation pattern in maize kernels. *BMC Plant Biology*, 17, 163.

Yan, J. and Fine, J. (2004) Estimating equations for association structures. *Statistics in Medicine*, 23, 859–874.

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, J., Ji, Y. and Zhang, L. (2007) Extracting three-way gene interactions from microarray data. *Bioinformatics*, 23, 2903–2909.

SUPPORTING INFORMATION

Tables and Figures referenced in Sections 3 and 4 are available with this paper at the Biometrics website on Wiley Online Library. R code and example data are available at the Biometrics website on Wiley Online Library. R code for implementing ZENCO is also available at http://www.github.com/zheny714/ZENCO.

How to cite this article: Yang Z, Ho Y-Y. Modeling dynamic correlation in zero-inflated bivariate count data with applications to single-cell RNA sequencing data. *Biometrics*. 2022;78:766–776. https://doi.org/10.1111/biom.13457