Comparison of Regression Methods to Identify Differential Expression in RNA-Sequencing Count Data from the Serial Analysis of Gene Expression

Ivan Arreola and David Han

Department of Management Science and Statistics, University of Texas at San Antonio, TX

ABSTRACT

Comparative RNA-sequencing analysis for the Serial Analysis of Gene Expression (SAGE) can help identify changes in gene expression which are characteristic to human diseases. Since the RNA-sequencing experiment measures gene expressions in the form of counts, usually with a large degree of skewness, the analysis methods based on continuous probability distributions such as the normal distribution are generally inappropriate for modeling this type of data. Currently, the parametric regression techniques for solving this problem are based on the well-known discrete probability distributions such as Poisson and negative binomial. In order to overcome this modeling challenge with higher flexibilities to account for a wide range of dispersion levels, here we introduce an alternative Generalized Linear Model (GLM) based on the Conway–Maxwell-Poisson distribution, also known as COM-Poisson or CMP distribution. The CMP regression model generalizes the standard Poisson and negative binomial regressions, and it is suitable for fitting count data with varying degrees of over- and under-dispersions. Using simulated and real SAGE datasets, the performance of the proposed method is assessed in comparison to the Poisson- and negative binomial-based regression models.

***Keywords*:** Conway-Maxwell-Poisson Regression, Count Data, Generalized Linear Models, RNA-Sequencing, Serial Analysis of Gene Expression

INTRODUCTION

During the last decades, mRNA sequencing technology and other sequencing-based genomic experiments such as SAGE were able to uncover significant patterns between gene expressions and metabolic pathways. These sequence-based findings have furthered our understanding of cellular functions, which are characteristic to human diseases such as cancers. In particular, the information obtained from SAGE is similar to those obtained from microarray experiments but there are several distinguishable differences [1]. First, SAGE uses sequencing techniques as opposed to competitive microarray hybridization. Second, microarrays give continuous expression values while SAGE gives discrete expression values. Finally, SAGE provides information of all genes in a given sample whereas microarrays only give information on the genes that have been printed on the array.

Figure 1 below describes how the sequence spellings are obtained from an RNA-sequencing experiment. In the usual case, a sample containing messenger ribonucleic acids (mRNA) are randomly shattered into small fragments and reverse transcribed into complimentary deoxyribonucleic acids (cDNA) [1-3]. The cDNA library is then amplified through a PCR machine and sequenced by using the Sanger method, resulting in thousands of short sequences known as “tags” or “reads.” The list of tags is then counted to tabulate the frequency of occurrences for each gene across a library.

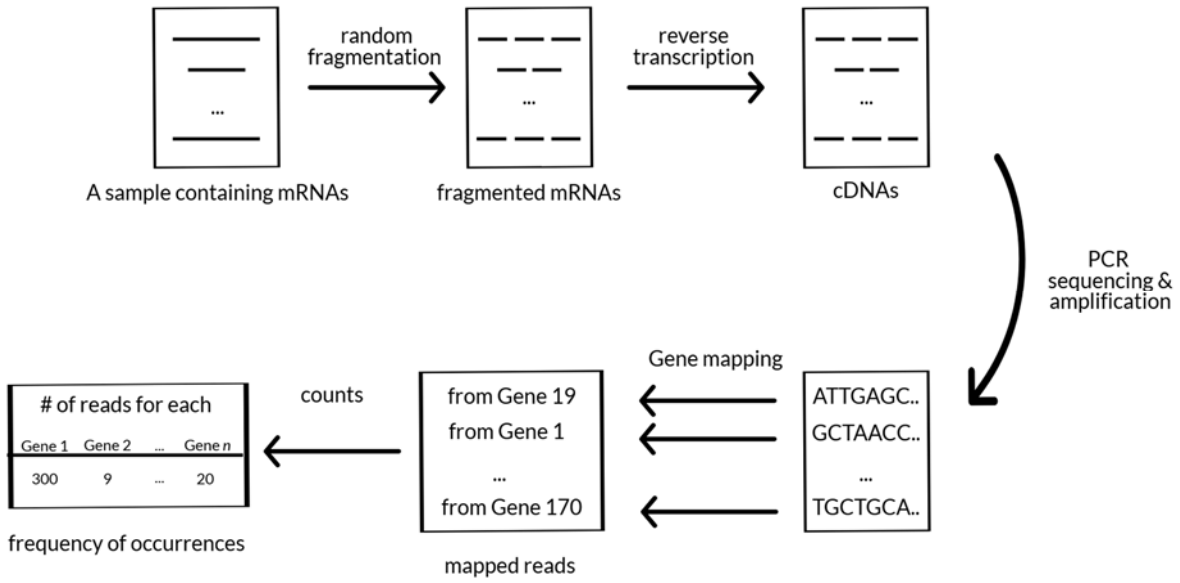

Figure 1. Flowchart for a typical RNA sequencing experiment [1]

RNA-sequencing experiments, just like microarrays, use comparative analysis. The experiments can come in the following forms: two classes, normal vs. tumor; multiple classes, primary tumor A vs. primary tumor B vs. primary tumor C vs. cell lines; quantitative such as continuously measured viral concentrations in a patient's blood specimen [2,3]. The goal of the comparative analysis is to identify differentially expressed genes that may be involved in the underlying biological functions of a cell.

Since the results of the RNA-sequencing experiments such as SAGE provide data in the form of the frequency of occurrences for each gene, their analyses require the statistical techniques for modeling discrete count data. This includes the popular Poisson, negative binomial, and logistic regressions. However, certain limitations exist in these methods. The Poisson regression assumes that the mean and variance are equal and this limits the manipulation of the underlying distribution [4,5]. On the other hand, the negative binomial regression only accounts for over-dispersion (where the variance is greater than the mean) but not under-dispersion. Lastly, the logistic regression model is useful but can be influenced by outliers. To overcome these obstacles, the CMP regression is suggested as it is able to account for various levels of dispersion in modeling discrete count data. As a member of the exponential family, the CMP distribution generalizes the Poisson and logistic models as well. Here we report the performance of the proposed CMP regression to analyze the SAGE count data in comparison to the Poisson, negative binomial, and logistic regression-based methods.

PROBABILITY MODELS & METHOD

Overdispersed Logistic Regression Over-dispersed logistic regression has been applied to model highly skewed count data [1]. For the change of expressions for a single tag, let us denote the set of counts $\{Y_i\}$ and the set of library sizes $\{n_i\}$, where i represents the specific library. Often, we have a covariate X_i describing the properties of a library. The

most common situation is to compare between normal and cancer groups. Then, the covariate X_i decides which group library i belongs to. The logistic model for proportions is expressed as

$$log\left(\frac{p_i}{1-p_i}\right) = logit(p_i) = \beta_0 + \beta_1 x_i$$

$$V(Y_i) = n_i p_i (1-p_i)$$

where x_i is 0 for the control group and 1 for the treatment group. The logistic framework is able to define what is to be modeled concerning the covariates and with accuracy of each measurement.

When over-dispersion is present, a quasi-likelihood and a hierarchical model are applied to handle the skewness. First, the quasi-likelihood inflates the variance for each observation by like amounts. With the scale term, σ_{QL}^2, we have the variance defined as

$$V(Y_i) = n_i p_i (1-p_i)\sigma_{QL}^2$$

Second, the hierarchical model is able to fit the proportions of the covariates from a positive distribution, which is implemented in an R package dispmod. With an estimation parameter ϕ, we have the variance of the beta-binomial model defined as

$$V(Y_i) = n_i p_i (1-p_i)[1 + (n_i - 1)\phi]$$

Under the logistic framework, analyzing differential expressions diminishes down to whether the regression coefficients are different from zero. When high skewness is present in count data, the logistic regression model is able to handle the outliers but can be easily influenced by extreme outliers, therefore resulting in false conclusions.

Poisson, Negative Binomial, and CMP Regressions A widely applied model for the count analysis is Poisson distribution, whose probability mass function is defined by

$$P(Y_i = y_i|x_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 0,1,2,\ldots$$

given a vector of covariates x_i and the mean parameter follows the log linear relationship defined by

$$log\ \mu_i = \beta_0 + \beta_1 x_i$$

When using the Poisson distribution for count analysis, the mean and variance must equal to each other. That is

$$V(Y_i|x_i) = E(Y_i|x_i) = \mu_i$$

This condition causes the Poisson regression to be restrictive, not allowing to capture over-dispersion in real datasets. Accommodating the over-dispersion can certainly improve the performance of the predictive capability of the model.

One distribution that allows for over-dispersion is the negative binomial distribution. The negative binomial distribution is derived from a mixture of both the gamma and Poisson random variates [5]. Its mean and variance are

$$E(Y_i|x_i) = \mu_i = e^{\beta_0 + \beta_1 x_i}$$

$$V(Y_i|x_i) = \mu_i\left[1 + \frac{1}{\theta}\mu_i\right] > E(Y_i|x_i)$$

Notice that the variance exceeds the mean. Making the substitution of $\alpha = \frac{1}{\theta}$, $\alpha > 0$, the negative binomial distribution can be reparametrized as

$$f(y_i|x_i) = \frac{\Gamma(y_i + \alpha^{-1})}{y_i!\,\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}$$

for $y_i = 0,1,2,\ldots$. The mean parameter again follows the log linear relationship defined by

$$log\ \mu_i = \beta_0 + \beta_1 x_i$$

As a special case, when α approaches 0, the negative binomial GLM converges to the Poisson GLM. Although the negative binomial distribution is able to handle over-dispersion, it is not in the case of under-dispersion. A flexible distribution that can model both under-dispersed and over-dispersed data will be able to shed more biological insight.

The CMP distribution is a generalization of the Poisson and negative binomial distributions and it allows for under-dispersion as well as over-dispersion [4]. The CMP distribution is defined as

$$P(Y_i = y_i|x_i) = \frac{1}{Z(\lambda_i, v_i)} \frac{\lambda_i^{y_i}}{(y_i!)^{v_i}}$$

where $y_i = 0,1,2, \ldots$ and the normalization factor is

$$Z(\lambda_i, v_i) = \sum_{n=0}^{\infty} \frac{\lambda_i^n}{(n!)^{v_i}}$$

with the regression models introduced by

$$log\ \lambda_i = \beta_0 + \beta_1 x_i$$

$$v_i = e^{(g_i'\delta)}$$

The corresponding mean and variance are

$$E[Y] = \frac{1}{Z(\lambda, v)} \sum_{j=0}^{\infty} \frac{j\lambda^j}{(j!)^v}$$

$$V[Y] = \frac{1}{Z(\lambda, v)} \sum_{j=0}^{\infty} \frac{j^2\lambda^j}{(j!)^v} - E[Y]^2$$

An additional parameter v provides the flexibility in modeling the tail behavior of the CMP distribution [5]. If, for instance, $v = 1$, the rate of decay is equal to that of the Poisson distribution. If $0 < v < 1$, then the rate of decay decreases, allowing to manipulate the model to have longer tails than the Poisson distribution (*i.e.*, over-dispersion). Lastly, if $v > 1$, then the rate of decay increases in a nonlinear form, therefore shortening the tail of the distribution and allowing for under-dispersed data. Hence, the CMP distribution possesses many advantages to model various levels of dispersion and generalizes the widely applied Poisson and negative binomial distributions.

A re-parameterization of the CMP model was proposed in order to provide a measure of central tendency that can be interpreted in the context of GLM [7]. By substituting $\lambda_i = {\mu_i}^{v_i}$, this formulation is written as

$$P(Y_i = y_i|x_i) = \frac{1}{S(\mu_i, v_i)} \left(\frac{\mu_i^{y_i}}{y_i!}\right)^{v_i}$$

where the new normalization factor is defined as

$$S(\mu_i, v_i) = \sum_{n=0}^{\infty} \left(\frac{\mu_i^n}{n!}\right)^{v_i}$$

with the regression model redefined as

$$log\ \mu_i = \beta_0 + \beta_1 x_i$$

The corresponding mean and variance are approximated by

$$E[Y] \approx \mu + \frac{1}{2}v - \frac{1}{2}$$

$$V[Y] \approx \frac{\mu}{v}$$

The dispersion is estimated as $\frac{V[Y]}{E[Y]} \approx \frac{1}{v}$. Likewise, $0 < v < 1$ for the over-dispersed data, $v = 1$ for the Poisson data, and $v > 1$ for the under-dispersed data.

In the next section, we report the performance of these regression models based on a simulation study. It was found that

in some cases the (reparametrized) CMP regression outperforms other GLM considered in this study (logistic regression, Poisson regression, and negative binomial regression).

SIMULATION STUDY

Random variates were simulated from Poisson, negative binomial, and CMP distributions for two classes: n_c control and n_t treatment groups. The selected coefficient values are $\beta_0 = 5$ and $\beta_1 = -1$, where β_0 is the intercept and β_1 is the slope. Furthermore, the covariate x_i is defined as $x_c = 0$ for control and $x_t = 1$ for treatment. The regression equation for the Poisson model is

$$log\ \lambda_p = \beta_0 + \beta_1 x_i$$

For the negative binomial model, it is

$$log\ \mu_{nb} = \beta_0 + \beta_1 x_i$$

For the CMP regression model, it is

$$log\ \mu_{cmp} = \beta_0 + \beta_1 x_i$$

All the measurements were applied to the Poisson, negative binomial, and (reparametrized) CMP regression models to test which distribution is more effective in finding differentially expressed sequences (*viz.*, statistical significance of β_1).

First, 8 Poisson random variables were generated with $n_c = 4$ for control and $n_t = 4$ for treatment with covariates $x_c = 0$ and $x_t = 1$ (Scenario A). Then, 8 negative binomial variables were generated with $n_c = 4$ and $n_t = 4$ with covariates $x_c = 0$ and $x_t = 1$. Two cases were considered in this situation: one simulation that converges to the Poisson distribution (Scenario B) and another that presents the usual over-dispersion (Scenario C). Lastly, 8 CMP random variables were generated with $n_c = 4$ and $n_t = 4$ with covariates $x_c = 0$ and $x_t = 1$. Similarly, we considered two cases, just as in the negative binomial situation: one simulation that converges to the Poisson distribution (Scenario D) and another that presents the over-dispersion (Scenario E). For analyzing these datasets, the following regression models were considered:

(1) Poisson GLM

(2) Poisson GLM with quasi-likelihood

(3) Poisson GLM with hierarchical model

(4) Negative binomial GLM

(5) (Reparametrized) CMP GLM

All the random variates were generated from the statistical software R and the analysis for all GLM-based methods were carried out in R and SAS.

Table 2 below presents the average *p*-values from the five regression models for each of the five scenarios aforementioned with 30 repetitions. Table 3 presents the average values of the log-likelihood and the Akaike Information Criterion (AIC) from the simulation study. The smaller the magnitudes of the log-likelihood are, the more significant the results are, and similarly for the AIC, the smaller the values are, the more significant the results are. From Table 2, it is observed that all the regression methods can handle the data from the exact or asymptotic Poisson distribution ($p < .0001$); see Scenarios A, B, and D. However, when dealing with the over-dispersed data from the negative binomial or CMP distribution, the CMP regression is able to do a better job; see Scenarios C and E. From Table 3, it is observed that the values of the log-likelihood are close to each other but overall the proposed CMP regression performs better than the Poisson and negative binomial GLMs. Also, the CMP GLM is shown to be better in handling the over-

Scenario A

Average p-values from Poisson random values with 30 repetitions

Model	β_0	β_1
(1)	.0001	.0001
(2)	.0001	.0001
(3)	.0001	.0001
(4)	.0001	.0001
(5)	.0001	.0001

Scenario B (Asymptotic Case)

Average p-values from the negative binomial random values with 30 repetitions

Model	β_0	β_1
(1)	.0001	.0001
(2)	.0001	.0001
(3)	.0001	.0001
(4)	.0001	.0001
(5)	.0001	.0001

Scenario C

Average p-values from the negative binomial random values with 30 repetitions

Model	β_0	β_1
(1)	.0001	.9900
(2)	.0024	.9980
(3)	.0001	.9980
(4)	.0005	.9985
(5)	.3788	.1705

Scenario D (Asymptotic Case)

Average p-values from the CMP random values with 30 repetitions

Model	β_0	β_1
(1)	.0001	.0001
(2)	.0001	.0001
(3)	.0001	.0001
(4)	.0001	.0001
(5)	.0001	.0001

Scenario E

Average p-values from the CMP random values with 30 repetitions

Model	β_0	β_1
(1)	.0001	.8092
(2)	.0001	.9542
(3)	.0001	.9523
(4)	.0001	.9517
(5)	.0001	.3800

Table 2. Average *p*-values from the simulation study of 30 repetitions

Scenario A

Average log-likelihood and AIC from Poisson random values

Model	log-likelihood	AIC
(3)	-20.505	45.009
(4)	-20.382	45.698
(5)	-19.734	45.468

Scenario B (Asymptotic Case)

Average log-likelihood and AIC from negative binomial random values

Model	log-likelihood	AIC
(3)	-20.371	44.741
(4)	-20.323	45.512
(5)	-19.839	45.678

Scenario C

Average log-likelihood and AIC from negative binomial random values

Model	log-likelihood	AIC
(3)	-80.685	165.371
(4)	-26.148	58.296
(5)	-26.821	59.643

Scenario D (Asymptotic Case)

Average log-likelihood and AIC from CMP random values

Model	log-likelihood	AIC
(3)	-20.684	45.369
(4)	-20.535	46.269
(5)	-19.851	45.703

Scenario E

Average log-likelihood and AIC from CMP random values

Model	log-likelihood	AIC
(3)	-85.202	174.404
(4)	-35.886	77.771
(5)	-35.836	77.672

Table 3. Average values of the log-likelihood and AIC from the simulation study

dispersed data from the CMP population; see Scenario E.

APPLICATION TO RNA SEQUENCES

Based on the results of the simulation study, we also assessed the proposed method in detecting differentially expressed genes in a real SAGE dataset, composed of the counts for the following three mRNA sequences: ATTTGAGAAG, TGCTGCCTGT, and GCGAAACCCT in 8 libraries, which include two normal colons (NC1 and NC2), two primary tumors (TU98 and TU102), and four cell lines (CACO2, HCT116, RKO, and SW837) [1]. The counts for tags and library sizes are also provided in [1]. The initial focus is on comparing the counts of tag ATTTGAGAAG between two libraries for normal colon and primary tumors. The following GLM models were used for this purpose.

(1) Logistic regression GLM
(2) Logistic regression GLM with quasi-likelihood
(3) Logistic regression GLM with hierarchical model
(4) Poisson GLM
(5) Poisson GLM with quasi-likelihood
(6) Poisson GLM with hierarchical model
(7) Negative binomial GLM
(8) (Reparametrized) CMP GLM

From Table 4, tag ATTTGAGAAG is found to be a significant sequence in GLM (1), (4), and (5). The significance of the logistic GLM agrees with the results in [1]. Next, the tag TGCTGCCTGT was examined and found interesting in [6]. It was first attempted to model the counts from two libraries, normal and tumor, ignoring the cell lines (two groups). Then, the counts from normal, tumor, and cell lines were modeled. Here we used two covariate vectors, $x_1 =$ $(0,0,1,1,0,0,0,0)$ and $x_2 = (0,0,0,0,1,1,1,1)$, each for two groups so that the modeling of normal and tumor does not intertwine with cell lines. The results are shown in Table 5.

Table 4.
Average *p*-values for tag ATTTGAGAAG

Model	β_0	β_1
(1)	.0001	.0001
(2)	.0001	.187
(3)	.00004	.186
(4)	.0001	.0001
(5)	.0001	.0001
(6)	.00004	.186
(7)	.00004	.230
(8)	.0001	.139

Table 5.
Average *p*-values for tag TGCTGCCTGT

	Two Groups		Three Groups		
Model	β_0	β_1	β_0	β_1	β_2
(3)	.038	.378	.007	.361	.299
(6)	.630	.200	.584	.099	.054
(7)	.623	.186	.583	.097	.053
(8)	.912	.908	.651	.621	.615

Similar to the case of tag ATTTGAGAAG, it was found that the logistic regression GLM agrees well with the results in [1]. Moreover, the models (6) and (7) for three groups were able to detect differential expressions, showing that tag TGCTGCCTGT might be characteristic to human diseases.

In the above case, we only compared the libraries of two groups, normal and tumor. Since it is more practical to compare three libraries (normal colon, primary tumor, and cell lines), we performed this analysis on tag GCGAAACCCT to examine the

capability of various GLM to detect differential expressions across multiple groups. With the addition of a new covariate, we could also create two cases, hypothetical covariate and hypothetical biomarker. The results are shown in Table 6 below.

Table 6.
Average *p*-values for tag GCGAAACCCT

Model	Hypothetical Covariate			Hypothetical Biomarker			
	β_0	β_1	β_2	β_0	β_1	β_2	β_3
(3)	.0001	.088	.082	.0060	.120	.113	.459
(6)	.0001	.119	.019	.0001	.185	.042	.982
(7)	.0001	.050	.006	.0002	.070	.011	.937
(8)	.0001	.023	.045	.0001	.0002	.0001	.020

In this particular case, the model (8) – CMP GLM – was able to identify the tag as differentially expressed for hypothetical covariate and biomarker. Additionally, the negative binomial GLM in both cases was able to identify the tag as differentially expressed.

Now, we examine the averages of the log-likelihood and AIC for all three tags: ATTTGAGAAG, GCGAAACCCT, and TGCTGCCTGT. From Table 7 below, the average log-likelihoods of tag ATTTGAGAAG for the negative binomial (7) and CMP (8) regressions are shown to be marginally close to each other but substantially better than the Poisson GLM with hierarchical model (6). However, in terms of AIC, the CMP GLM is stands out to be clearly a better model.

Similarly, Table 8 shows the averages of the log-likelihood and AIC for tag TGCTGCCTGT. In two groups, both the negative binomial and CMP are exceptionally close in terms of the log-likelihood and AIC. In three groups, the same observation could be made but the CMP GLM was found to be slightly better than the negative binomial results. Finally, from Table 9 for tag GCGAAACCCT, similar observations could be made just like in Table 8. For the log-likelihood values and AIC of the hypothetical covariate case, both the negative binomial and CMP GLM are again close to each other but the negative binomial GLM was found to be slightly better than the CMP one this time. On the other hand, in the hypothetical biomarker case, the CMP GLM was found to be slightly better than the negative binomial regression.

Table 7.
Averages log-likelihood and AIC
for tag ATTTGAGAAG

Model	log-likelihood	AIC
(6)	-540.660	109.89
(7)	-51.944	1085.00
(8)	-52.289	110.58

Table 8.
Averages log-likelihood and AIC
for tag TGCTGCCTGT

Model	Two Groups		Three Groups	
	log-likelihood	AIC	log-likelihood	AIC
(6)	-12.321	28.321	-35.897	77.794
(7)	-8.182	22.365	-22.300	52.600
(8)	-8.182	22.364	-22.221	52.442

Table 9.
Averages log-likelihood and AIC
for tag GCGAAACCCT

Model	Hypothetical Covariate		Hypothetical Biomarker	
	log-likelihood	AIC	log-likelihood	AIC
(6)	-169.601	345.203	-109.561	227.122
(7)	-40.177	88.354	-40.175	90.349
(8)	-41.887	91.773	-39.551	89.102

Overall, the CMP regression is suggested as a competitive or better model to handle over-

dispersed data as well as the Poisson data and under-dispersed data.

SUMMARY

The RNA-sequencing experiments such as SAGE produce discrete data in the form of the frequency of gene tags/reads. Because of this, their analyses require the count data regression techniques, which differ from the conventional continuous data regressions. This work recommends the CMP regression model as a better alternative to the classical Poisson and negative binomial regressions because of its versatility to accommodate all levels of dispersion in modeling discrete count data. It also includes the classical models as special asymptotic cases. The performance results from the simulation study and the SAGE application demonstrate the utility of the CMP regression for detecting statistically significant differentially expressed genes for biomedical research. Further research in this area requires a sound development of multiple-comparison procedures when processing a large number of tags with the false discovery rates under control.

REFERENCES

[1] Baggerly, A. K., Deng, L., Morris, S. J., and Aldaz, M. C. (2004). "Overdispersed logistic regression for SAGE: Modeling multiple groups and covariates." *BMC Bioinformatics,* **5**: 144.

[2] Li, J., Witten, M. D., Johnstone, M. I., and Tibshirani R. (2012). "Normalization, testing, and false discovery rate estimation for RNA-sequencing data." *Biostatistics,* **13:** 523–538.

[3] Li, J. and Tibshirani, R. (2011). "Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data." *Statistical Methods in Medical Research,* **22:** 519–536.

[4] Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). "A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution." *Journal of the Royal Statistical Society – Series C,* **54:** 127–142.

[5] The COUNTREG Procedure, *SAS/ ETS 13.2 User Guide.* *https://support.sas.com/documentation/onlinedoc/ets/132/countreg.pdf*

[6] Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruben, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. (1997). "Gene expression profiles in normal and cancer cells." *Science,* **276:** 1268–1272.

[7] Guikema, S. D. and Coffelt, J. P. (2008). "A flexible count data regression model for risk analysis." *Risk Analysis,* **28:** 213–223.