PLOS ONE

RESEARCH ARTICLE

Penalized negative binomial models for modeling an overdispersed count outcome with a high-dimensional predictor space: Application predicting micronuclei frequency

Rebecca R. Lehman[1☯], **Kellie J. Archer**[2☯*]

1 United Network for Organ Sharing, Richmond, VA, United States of America, **2** Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, United States of America

☯ These authors contributed equally to this work.
* archer.43@osu.edu

Citation: Lehman RR, Archer KJ (2019) Penalized negative binomial models for modeling an overdispersed count outcome with a high-dimensional predictor space: Application predicting micronuclei frequency. PLoS ONE 14(1): e0209923. https://doi.org/10.1371/journal.pone.0209923

Editor: Fabio Rapallo, Universita degli Studi del Piemonte Orientale Amedeo Avogadro, ITALY

Received: July 16, 2018

Accepted: December 13, 2018

Published: January 8, 2019

Data Availability Statement: Data were downloaded from Gene Expression Omnibus, accession number GSE31836.

Funding: Research reported in this publication was supported by the National Library of Medicine (https://www.nlm.nih.gov) of the National Institutes of Health under Award Number R01 LM011169 & the National Institute of Environmental Health Sciences (https://www.niehs.nih.gov) under Award Number T32 ES007334. The content is solely the

Abstract

Chromosomal aberrations, such as micronuclei (MN), have served as biomarkers of genotoxic exposure and cancer risk. Guidelines for the process of scoring MN have been presented by the HUman MicroNucleus (HUMN) project. However, these guidelines were developed for assay performance but do not address how to statistically analyze the data generated by the assay. This has led to the application of various statistical methods that may render different interpretations and conclusions. By combining MN with data from other high-throughput genomic technologies such as gene expression microarray data, we may elucidate molecular features involved in micronucleation. Traditional methods that can model discrete (synonymously, count) data, such as MN frequency, require that the number of explanatory variables (*P*) is less than the number of samples (*N*). Due to this limitation, penalized models have been developed to enable model fitting for such over-parameterized datasets. Because penalized models in the discrete response setting are lacking, particularly when the count outcome is over-dispersed, herein we present our penalized negative binomial regression model that can be fit when $P > N$. Using simulation studies we demonstrate the performance of our method in comparison to commonly used penalized Poisson models when the outcome is over-dispersed and applied it to MN frequency and gene expression data collected as part of the Norwegian Mother and Child Cohort Study. Our `countgmifs` R package is available for download from the Comprehensive R Archive Network and can be applied to datasets having a discrete outcome that is either Poisson or negative binomial distributed and a high-dimensional covariate space.

Introduction

More than 85% of all cancers are associated with acquired chromosomal or genetic alterations. Various cytogenetic endpoints have been used for cancer risk assessment, including structural

responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests: The authors have declared that no competing interests exist.

chromosomal aberrations, aneuploidy, while sister chromatid exchanges have been useful biomarkers of exposure [1]. Micronuclei (MN) have also been used both for cancer risk assessment and to assess exposure to genotoxic agents. Micronuclei (MN) are formed in dividing cells from either whole chromosomes or chromosome fragments that lag behind [2], do not attach to the mitotic spindle prior to cytokinesis, so that after cell division, they appear to be small extranuclear bodies (Fig 1). The presence of MN, particularly at high frequencies, is taken to reflect chromosomal abnormalities. Various investigators have reported higher MN frequencies among people who were exposed to a toxic agent compared to controls not exposed [3–6]. Also, previous research has shown that higher MN frequencies are associated with a higher risk of cancer development [7, 8] and that MN frequencies are higher in subjects with cancer compared to those without cancer [9–14]. As a specific example, MN formation was increased due to genotoxic hepatocarcinogen exposure compared to non-genotoxic exposure [15], indicating MN formation is due to chromosomal damage. Thus, MN serve as an early marker for carcinogenesis. In fact, MN frequency has been used in biomonitoring [16, 17], occupational exposure [3, 4, 18–20], environmental exposure [5, 6, 21–24], and as previously mentioned, in cancer research studies [7, 9–14, 25–31]. MN frequency is also relevant in other developmental, age-related, degenerative diseases (e. g. Alzheimer's disease, Parkinson's disease) [6]. Therefore, because MN are objectively measured they are a useful biomarker for assessing both genotoxic exposure and cancer risk and may serve as an indicator of a pathogenic process.

The Cytokinesis block micronucleus assay (CBMN) assay is commonly used to score MN. Briefly, the CBMN assay uses cytochalasin-B, which stops cells from performing cytokinesis but does not stop nuclear division, giving rise to cells that are binucleated [2, 8]. Guidelines for scoring MNs have been established by the HUman MicroNucleus (HUMN) project to minimize inter-rater variation [6, 32]. MN frequency is generally reported as the number of binucleated cells containing at least one MN. Therefore, MN frequency is a discrete or count variable. Other unique nuclear anomalies detectable using the CBMN/cytome assay include nucleoplasmic bridges and nuclear buds, which seem to be caused by different mechanisms in comparison to MN [33]. In fact, buds, also called broken eggs, are considered to be a marker of gene amplification, being conjectured to arise due to errors in replication during the S phase of the cell cycle [32, 34]. Thus scoring each of these anomalies provides a comprehensive assessment of genotoxic exposure and genetic damage [35, 36].

Although the HUMN project developed guidelines for assay performance, they have not addressed how to statistically analyze data generated by the assay. Ceppi et al [37] reviewed 63 studies that analyzed MN frequency in buccal cells with respect to their study design and analytical methods applied. Most frequently the studies involved two-group comparisons so that the t-test, the non-parametric Mann-Whitney U-test were most frequently applied (38.1%, 31.7% of studies, respectively). Although the non-parametric tests applied do not require an underlying distributional assumption, they are unable to adjust for confounding factors and often result in a loss of statistical power. While linear regression, ANOVA, and ANCOVA can be used to adjust for confounding variables and effect modifiers, problematically MN frequency rarely follows a Gaussian distribution (Fig 2). As mentioned, MN frequency is a discrete count outcome. Discrete probability distributions differ from those for continuous (takes on an infinite number of possible values), nominal (categorical variable), and ordinal ($\geq$ 3-level categorical variable where the categories have a natural ordering imposed, such as small, medium, large) variables. When MN frequency is over-dispersed, that is, the distribution is not well-described by the mean and variance having the same parameter, the discrete negative binomial (NB) distribution fits better than the Poisson distribution, while the

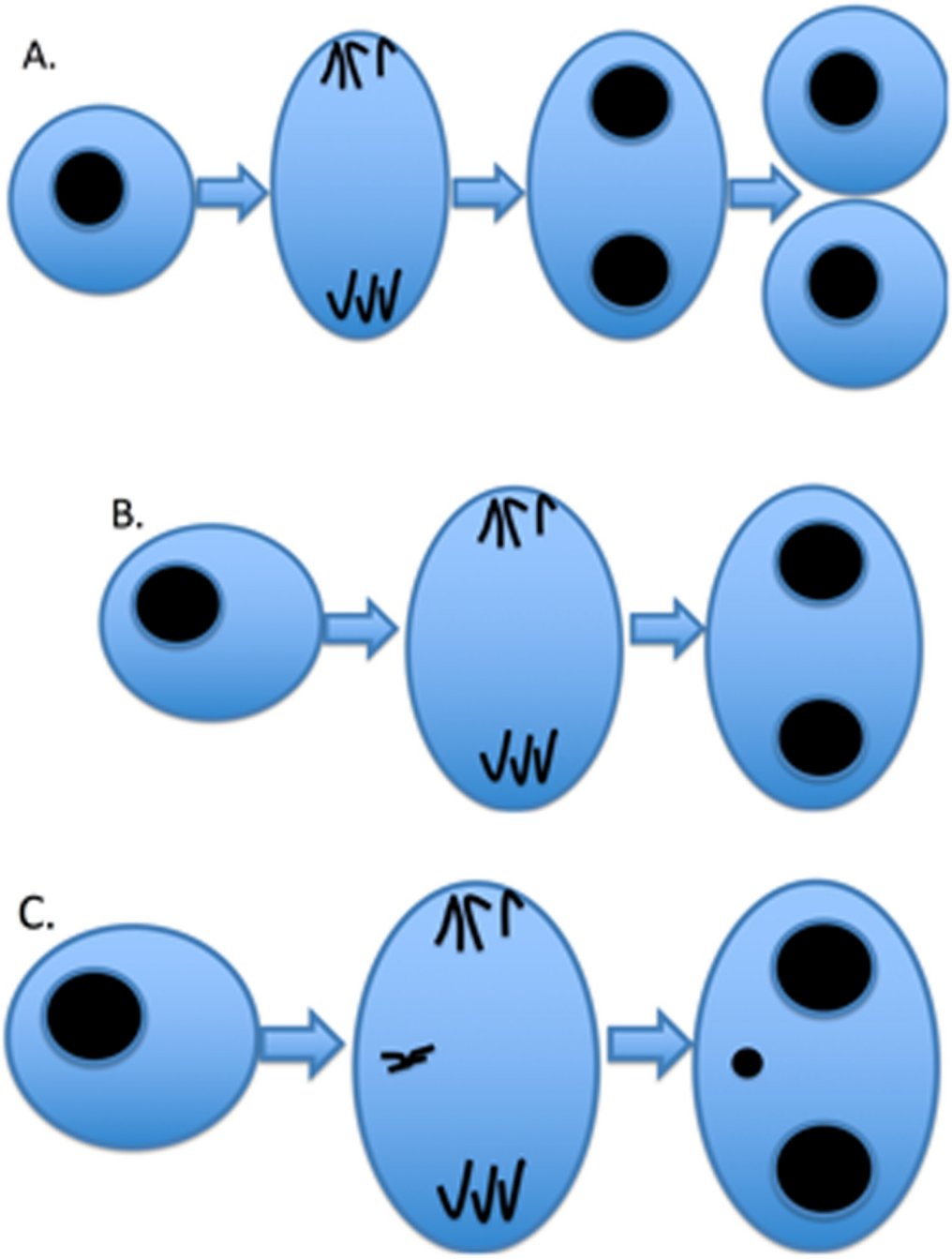

Fig 1. CBMN/Cytome assay. A: Typical process of cell division. B: Application of cytochalasin-B prevents cytokinesis to give rise to binucleated cells. C: Cell treated with cytochalasin-B that contains a whole chromosome lagging behind that does not attach to the mitotic spindle. The small extranuclear body in the binucleated cell is a MN.

https://doi.org/10.1371/journal.pone.0209923.g001

Gaussian distribution poorly fits the MN data (Fig 2). Over-dispersion may be due to the data including more zero count responses than expected, positive correlation between the count responses, or due to correlation between an explanatory variable and the error term [38]. The skewed distribution of MN frequency demonstrates excess zeros, and suggests that methods

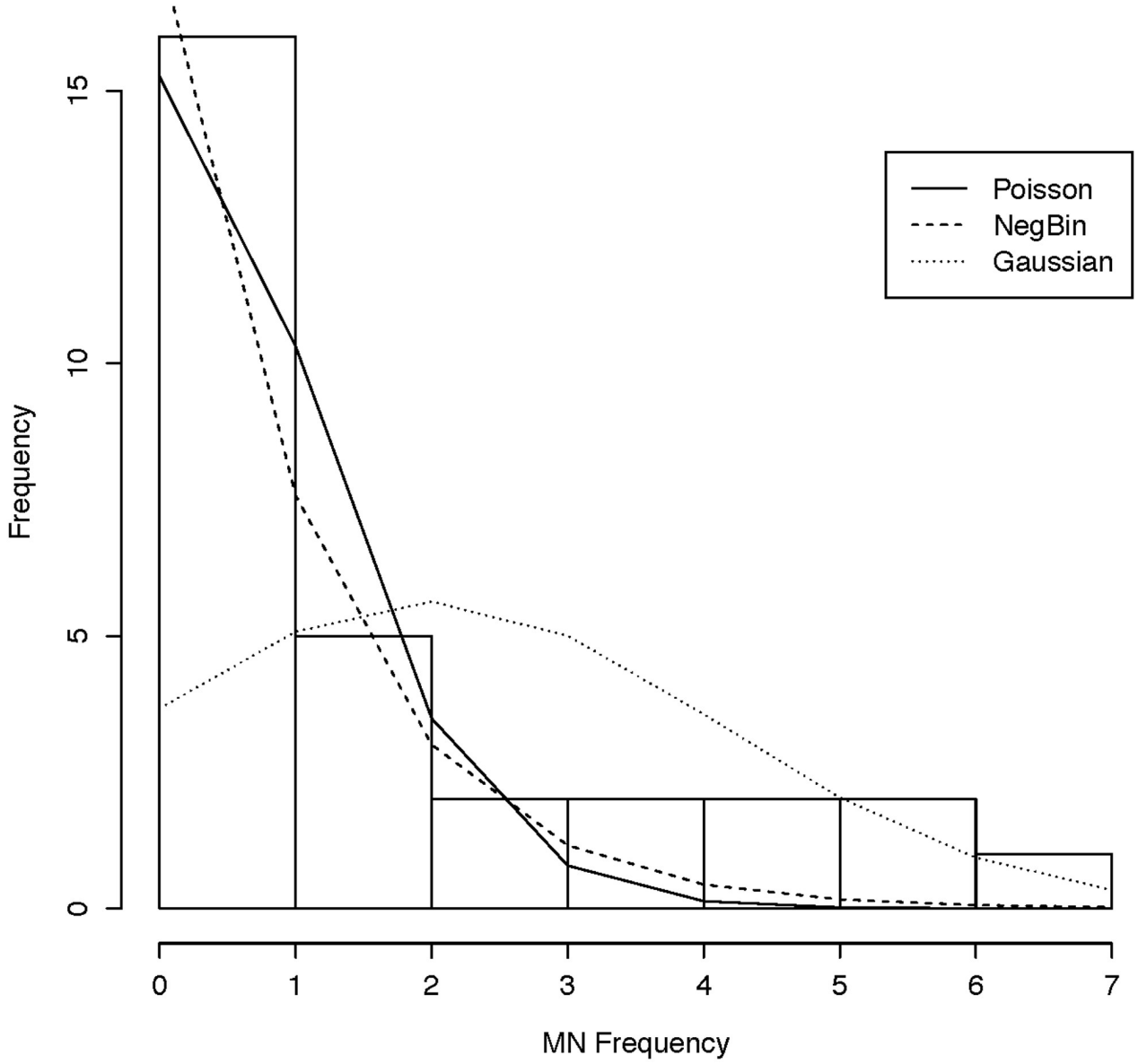

Fig 2. MN frequency distribution. Histogram of MN frequency from the cheek opposite an oral carcinoma with lines representing the Poisson (solid), Negative Binomial (dashed), and Gaussian (dotted) distributions. Data from Ramirez & Saldanha [39].

https://doi.org/10.1371/journal.pone.0209923.g002

based on the Gaussian distribution may lead to inappropriate inferences. Unfortunately, authors have not routinely reported whether basic assumptions of the inferential tests applied were met [37].

Among the 63 studies reviewed [37], non-Gaussian multivariable models applied included logistic ($N = 1$), log-linear ($N = 1$), Poisson ($N = 1$), and NB regression ($N = 4$). The logistic and log-linear models only model categorical data (e.g., MN are present versus absent) whereas Poisson and NB directly model count data. Because MN frequency is generally reported as the number of binucleated cells containing at least one MN, accounting for the number of binucleated cells scored is important. Therefore, advantages of the Poisson and NB models include

that they: (1) are well-suited for modeling skewed count distributions; (2) can adjust for confounding variables such as age, gender, smoking status [37]; and (3) can account for the total number of cells scored per patient, which may vary from sample to sample. For these reasons, Ceppi et al [37] recommended using NB or Poisson regression models when analyzing MN data.

In the last decade, high-throughput genomic platforms have been increasingly used to identify molecular features associated with disease status, exposure, or outcome. Recently, some studies have collected both MN frequency and gene expression data to confirm the genotoxicity of the exposure studied and to assess the molecular impact of exposure, respectively. Unfortunately, the association between gene expression and MN frequency was often not explored [15, 40, 41]. We know that many disease-causing mutations do not have complete penetrance, and penetrance is affected by environmental factors, age, epigenetics, and other genetic modifiers. Moreover, conditions with a complex inheritance pattern, such as cancer, are likely polygenic and have environmental contributions.

As previously mentioned, MN frequency is a discrete count outcome. Though Poisson and negative binomial regression can model a discrete response, they traditionally require that the number of covariates (P) is less than the number of samples (N) [38]. However, in high-throughput genomic datasets, $P > N$, often by several orders of magnitude. While methods such as principle components analysis have been used to reduce the total number of variables in the dataset, thus permitting model fitting, the resulting features are complicated linear combinations of genomic features, making downstream gene interpretations difficult. As an alternative, the least absolute shrinkage and selection operator (LASSO) [42, 43], also referred to as L1 penalized models, can be used to fit continuous [44, 45], binary [46, 47], and survival [48–51] models when $P > N$. Downstream gene interpretations are more straight-forward, because the original features are included in the model (though they are often first centered and scaled) [42, 43]. There are various computational approaches for estimating L1 penalized models including least angle regression (LARS) for the ordinary linear regression setting [52] and the predictor-corrector [53] and cyclical coordinate descent [54] algorithms for generalized linear models. Also, Hastie et al (2007) demonstrated that the solution path of the incremental forward stagewise (IFS) method is equivalent to the LASSO when the LASSO path is monotone for each coefficient and Efron et al (2004) described necessary and sufficient conditions for every coefficient path to be monotone. IFS was described in detail for ordinary linear regression but also for general convex loss functions, with logistic regression as a specific example [55]. For general convex loss functions, incremental updates are made to the coefficient having the smallest negative gradient of the log-likelihood at the current model estimates, hence the algorithm is called the generalized monotone incremental forward stagewise method. While the incremental nature of the algorithm naturally leads to longer computational time, the monotone paths are much smoother than those produced by other L1 algorithms which sometimes yield widely fluctuating paths, particularly when covariates are highly correlated. We have also recently extended the generalized monotone incremental forward stagewise (GMIFS) method [55] to high-dimensional ordinal response [56–61] and Poisson regression [62] settings. However, penalized methods have not been fully extended to discrete response setting when over-dispersion is present. We postulate that it is of interest to identify molecular features associated with micronucleation, as such features may elucidate important mechanisms in genotoxicity and carcinogenesis. Herein we describe a multivariable discrete response modeling method that can be applied to a high-dimensional covariate space when the count outcome is over-dispersed, such as the case for MN data.

Materials and methods

Poisson regression

When modeling the number of times an event occurs in either time or space, a generalized linear model (GLM) such as Poisson or negative binomial regression is commonly applied. Let $i = 1, \ldots, N$ be the number of observations, y_i represent a Poisson distributed random variable. Let the expected value of y_i be written as $\mathbb{E}(y_i) = \mu_i$. Then the conditional probability $P(y_i|\mu_i)$ for each observation i, subsequently the likelihood $L(\mu|\mathbf{y})$ are represented by

$$P(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \qquad \text{and} \qquad L(\mu|\mathbf{y}) = \prod_{i=1}^{N} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}. \tag{1}$$

Mathematically it is easier to maximize the log-likelihood which is given by

$$\ell(\mu|\mathbf{y}) = \sum_{i=1}^{N} (y_i \log \mu_i - \mu_i - \log(y_i!)). \tag{2}$$

Thus, we are looking for the value of μ that maximizes the log-likelihood in Eq 2. When fitting a GLM, a non-linear transformation, or link function, of the mean response is applied, which is a linear function of the covariates [38]. The link function for a Poisson regression model is $\log(\mu)$, thus, $\mu_i = \exp(\gamma_0 + \mathbf{x}_i^\top \boldsymbol{\beta})$ where γ_0 is the intercept and $\boldsymbol{\beta}$ is the vector of regression coefficients. Further, MN frequency is scored from a large number of binucleated cells, c. If the number of cells examined varies by subject, modeling MN frequency as a rate is more appropriate. We note that expressing a discrete response as a rate then transforming the rate to get it to adhere to a Gaussian distribution so that traditional linear models can be fit, does not properly account for the variation observed in the numerator and denominator terms. Therefore, Poisson and NB regression models that explicitly include the denominator term are more appropriate. The Poisson regression model for the expected count per unit of c_i, where c_i is the number of cells scored in our application, is $\mathbb{E}(y_i/c_i) = \mu_i = \exp(\gamma_0 + \mathbf{x}_i^\top \boldsymbol{\beta})$ which is equivalent to

$$\mu_i = \exp(\gamma_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \log(c_i)) \tag{3}$$

so that $\log(c_i)$ is an offset term. Traditionally count models are estimated by maximum likelihood or an iteratively re-weighted least squares algorithm [38]. For the rate-based model, the log-likelihood expressed with covariates is

$$\ell(\mu|\mathbf{y}) = \sum_{i=1}^{N} (y_i(\gamma_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \log(c_i)) - \exp(\gamma_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \log(c_i)) - \log(y_i!)). \tag{4}$$

Negative binomial regression

For the Poisson regression model, the distributional assumption is that the responses are from a Poisson distribution, which is a member of the exponential family. Members of the exponential family have a common characteristic that the variance of the response can be expressed as the product of a dispersion parameter, ϕ, and the variance function, $\mathrm{Var}(\mu_i)$. In other words, $\mathrm{Var}(Y_i) = \phi \mathrm{Var}(\mu_i)$. For the Poisson distribution, the dispersion parameter is a fixed constant ($\phi = 1$) and does not require estimation; therefore, $\mathrm{Var}(\mu_i) = \mu$.

Because the Poisson distribution has one parameter μ which represents both the mean and variance of the count outcome, the Poisson model is equi-dispersed. As previously mentioned, over-dispersion is present when the variance exceeds the mean, which is a common

occurrence in observed data. If not properly accounted for, over-dispersion results in incorrect standard errors and thus potentially incorrect inferences. The Pearson dispersion can be examined as a means for checking for over-dispersion in a Poisson regression model. If the Pearson dispersion exceeds 1.0, over-dispersion may be present. Alternative models such as the NB model, zero-inflated models, truncated models, or quantile count models can be used, though NB regression was found to best handle over-dispersed discrete response data [63]. Extensive work has been done with the Poisson and negative binomial distributions in the traditional statistical setting. However, there are limited methods for analyzing a count outcome with a high-dimensional predictor space.

Penalized Poisson regression

Development of high-dimensional methods for count response data have largely been restricted to the Poisson distribution [53, 54, 62, 64], especially for longitudinal settings [65–67]. The `glmpath` [53], `glmnet` [54], and `nnlasso` [64]R packages each provide an L_1 regularization path algorithm that seek either a LASSO solution, which minimizes $-\ell(\boldsymbol{\beta}; y) + \lambda \sum_{p=1}^{P} |\beta_p|$, or an elastic net solution by estimating coefficients over a grid of λ values, where $\lambda > 0$ is the regularization parameter. Basically, `glmpath` starts with the smallest value of the regularization parameter, λ_{max}, at which only the intercept is non-zero. Thereafter, a grid of λ values from λ_{max} to 0 is identified which allows other covariates to enter the model. The method uses the predictor-corrector algorithm to estimate the next value of λ that will change the current set of non-zero parameter estimates in the model (the active set) and then finds the solution to $\boldsymbol{\beta}$ at that value of λ. The `glmnet` algorithm also starts with the smallest value of the regularization parameter, λ_{max}, at which only the intercept is non-zero but then uses a sequence of 100 λ values that are evenly spaced on the log-scale. It uses cyclical coordinate descent to solve the elastic net penalized model, which minimizes $-\ell(\boldsymbol{\beta}; y, \alpha) + \lambda\left(\sum_{p=1}^{P}(\frac{1}{2}(1-\alpha)\beta_p^2 + \alpha|\beta_p|)\right)$ where α is a fixed proportion representing a compromise between the ridge and L_1 penalties. While the `nnlasso` algorithm also starts with the smallest value of the regularization parameter, λ_{max}, at which only the intercept is non-zero, it then uses a sequence of k λ values that are evenly spaced on a linear scale, and then uses a multiplicative iterative-Armijo algorithm for estimating model parameters under a non-negative constraint at each value of λ. While various penalized Poisson models can be fit to high-dimensional data, the focus of this paper is to extend the generalized monotone incremental forward stagewise (GMIFS) method to the negative binomial regression setting and demonstrate its effectiveness analyzing over-dispersed count outcome data.

Proposed NB GMIFS

The proven GMIFS technique from our previous studies [58, 60–62] will be extended to a new setting, the NB model. The NB model is derived from a Poisson-gamma mixture distribution [38], such that the variance is $\mu + \mu^2/\phi$. Unlike the Poisson model, ϕ must be estimated. Most often the parameter ϕ is expressed as its inverse, $\alpha = 1/\phi$, such that α is the heterogeneity parameter that directly models the amount of extra-dispersion in the data. Our proposed NB GMIFS method includes approaches for initializing the intercept, coefficients for the unpenalized subset of predictors, and heterogeneity parameter α; methods for updating these estimates after each iterative update of the penalized subset of predictor variables; derivatives for identifying which covariate to update; and convergence criteria. The NB probability mass

function is given by,

$$f(y;\mu,\alpha) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha}\left(1-\frac{1}{1+\alpha\mu_i}\right)^{y_i} \tag{5}$$

so the likelihood is

$$L(\mu;y,\alpha) = \prod_{i=1}^{N}\frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)}\left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha}\left(1-\frac{1}{1+\alpha\mu_i}\right)^{y_i} \tag{6}$$

which can be re-arranged as

$$\begin{aligned} L(\mu;y,\alpha) = \prod_{i=1}^{N}\exp\big(&\log\Gamma(y_i+1/\alpha) - \log\Gamma(y_i+1) - \log\Gamma(1/\alpha) + \\ &1/\alpha\log\left(\frac{1}{1+\alpha\mu_i}\right) + y_i\log\left(1-\frac{1}{1+\alpha\mu_i}\right)\big) \end{aligned} \tag{7}$$

Similar to Poisson regression, the log link function is used and when the outcome is a rate with the offset term c_i such that $\mathbb{E}(y_i/c_i) = \mu_i$ or $\mathbb{E}(y_i) = c_i\mu_i$. Therefore, the likelihood is

$$\begin{aligned} L(\mu;y,\alpha) = \prod_{i=1}^{N}\exp\big(&\log\Gamma(y_i+1/\alpha) - \log\Gamma(y_i+1) - \log\Gamma(1/\alpha) + \\ &1/\alpha\log\left(\frac{1}{1+\alpha\mu_i c_i}\right) + y_i\log\left(1-\frac{1}{1+\alpha\mu_i c_i}\right)\big) \end{aligned} \tag{8}$$

and the corresponding log-likelihood is

$$\begin{aligned} \ell(\mu;y,\alpha) = \sum_{i=1}^{N}\big(&\log\Gamma(y_i+1/\alpha) - \log\Gamma(y_i+1) - \log\Gamma(1/\alpha) + \\ &1/\alpha\log\left(\frac{1}{1+\alpha\mu_i c_i}\right) + y_i\log\left(1-\frac{1}{1+\alpha\mu_i c_i}\right)\big) \end{aligned} . \tag{9}$$

Given $\mathbb{E}(y_i/c_i) = \mu_i = \exp(\gamma_0 + \mathbf{x}_i^\top\boldsymbol{\beta})$ such that $\mathbb{E}(y_i) = c_i\exp(\gamma_0 + \mathbf{x}_i^\top\boldsymbol{\beta}) = \exp(\gamma_0 + \mathbf{x}_i^\top\boldsymbol{\beta} + \log(c_i))$, the log-likelihood with covariates reflected is

$$\begin{aligned} \ell(\boldsymbol{\beta};y,\alpha) = \sum_{i=1}^{N}\big(&\log\Gamma(y_i+1/\alpha) - \log\Gamma(y_i+1) - \log\Gamma(1/\alpha) \\ &+1/\alpha\log\left(\frac{1}{1+\alpha\exp(\gamma_0 + \mathbf{x}_i^\top\boldsymbol{\beta} + \log(c_i))}\right) \\ &+y_i\log\left(1-\frac{1}{1+\alpha\exp(\gamma_0 + \mathbf{x}_i^\top\boldsymbol{\beta} + \log(c_i))}\right)\big) \end{aligned} \tag{10}$$

The heterogeneity parameter in the NB model that accounts for the extra-dispersion can be selected *a priori* or estimated outside of the GLM framework then treated as a constant. Most NB modeling software employs iteratively re-weighted least squares, where α is estimated and its estimand is inserted into the equation as a constant and maximum likelihood estimation is applied. We examined the performance of three different methods (maximum likelihood,

equating the deviance to the residual degrees of freedom, and Hilbe's method of moments), for estimating the heterogeneity parameter using a small simulation study and found that Hilbe's method of moments estimator performed well. Therefore in our proposed GMIFS NB model, we estimate α using Hilbe's method of moments estimator prior to the iterative procedure [38]. Similar to our GMIFS Poisson approach, we partition the design matrix, $\mathbf{x}$, in Eq 10 into two parts, $\mathbf{x}_j$ and $\mathbf{x}_k$, where $j = 1, \ldots, J$ is the set of unpenalized predictors, $k = 1, \ldots, K$ is the set of penalized predictors and $J + K = P$. The unpenalized predictors are those that we wish to force into the model, such as gender, age and smoking status which researchers consider important predictors of MN frequency [37] and their values are in the $\mathbf{x}_{ij}$ design matrix for subject i. The thousands of features from a high-throughput genomic experiment are the penalized predictors for which we seek a parsimonious model and their values are in the $\mathbf{x}_{ik}$ design matrix for subject i. The parameter vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ correspond to the unpenalized subset and penalized subset of predictors, respectively.

The algorithm proceeds in an iterative fashion and updates one of the penalized covariates by a small incremental amount at each step. To determine which penalized covariate is to be updated at each step the gradient of the log-likelihood is used. Thus we need to calculate the first derivative of the log-likelihood corresponding to each penalized predictor. The log-likelihood written in terms of γ_0, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is

$$\begin{aligned}\ell(\boldsymbol{\beta}; y, \alpha) = \sum_{i=1}^{N} \Big(& \log\Gamma(y_i + 1/\alpha) - \log\Gamma(y_i + 1) - \log\Gamma(1/\alpha) \\ & + 1/\alpha \log\Big(\frac{1}{1 + \alpha \exp(\gamma_0 + \mathbf{x}_{ij}^{\top}\boldsymbol{\gamma} + \mathbf{x}_{ik}^{\top}\boldsymbol{\beta} + \log(c_i))}\Big) \\ & + y_i \log\Big(1 - \frac{1}{1 + \alpha \exp(\gamma_0 + \mathbf{x}_{ij}^{\top}\boldsymbol{\gamma} + \mathbf{x}_{ik}^{\top}\boldsymbol{\beta} + \log(c_i))}\Big)\Big)\end{aligned} \tag{11}$$

and the first derivative with respect to β written in terms of γ_0, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in matrix notation is

$$\frac{\partial \ell}{\partial \beta} = \mathbf{x}_{p_k}^{\top}(\mathbf{y} - \exp(\gamma_0 + \log(c_i) + \mathbf{x}_j^{\top}\boldsymbol{\gamma} + \mathbf{x}_k^{\top}\boldsymbol{\beta}))/(1 + \alpha \exp(\gamma_0 + \log(c_i) + \mathbf{x}_j^{\top}\boldsymbol{\gamma} + \mathbf{x}_k^{\top}\boldsymbol{\beta})). \tag{12}$$

Once we know which covariate to update, we need to determine the sign (+ or −) of the update. Rather than calculating the second derivative, an expanded covariate space can be used to get the direction of the update [55]. Using the previous notation for the unpenalized ($\mathbf{x}_j$) and penalized ($\mathbf{x}_k$) variables, the expanded covariate space is $\tilde{\mathbf{x}} = [\mathbf{x}_j : \mathbf{x}_k : -\mathbf{x}_k]$ where $[\mathbf{x}_k : -\mathbf{x}_k]$ have been standardized. For the penalized predictors in the $[\mathbf{x}_k : -\mathbf{x}_k]$ component, let $\beta_1, \ldots, \beta_K$ be the positive coefficient estimates and $\beta_{K+1}, \ldots, \beta_{2K}$ be the coefficient estimates of the negative version of $\mathbf{x}_k$. Our NB GMIFS algorithm using the expanded covariate set is

1. Initialize the components of $\hat{\boldsymbol{\beta}}^{(s)} = 0$ at step $s = 0$ and initialize α using Hilbe's method of moments.
2. Estimate the intercept γ_0 and the unpenalized coefficients γ_j where $j = 1, \ldots, J$ using a maximization algorithm of the log-likelihood.
3. Considering $\hat{\alpha}$, $\hat{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}$ fixed, find the predictor $\mathbf{x}_m$ where $m = \underset{2K}{\operatorname{argmin}}\left(-\frac{\partial \ell}{\partial \beta_k}\right)$ at the current estimate $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(s)}$.

4. Update the corresponding coefficient $\hat{\beta}_m^{(s+1)} \leftarrow \hat{\beta}_m^{(s)} + \epsilon$ to yield a new vector of parameter estimates.
5. Update γ_0 and the unpenalized coefficients, γ_j, by maximum likelihood considering the $\hat{\boldsymbol{\beta}}^{s+1}$ from step 4 as fixed.
6. Re-estimate α using Hilbe's method of moments method.
7. Repeat steps 2-6 until the difference between successive log-likelihoods is less than a pre-specified tolerance, τ.
8. The final coefficient estimates are obtained by subtracting the pairs, $\beta_1 - \beta_{K+1}, \ldots, \beta_K - \beta_{2K}$, to yield $\hat{\boldsymbol{\beta}}$.

Hastie et al (2007) [55] did not specify a stopping criteria but recommended to repeat the steps many times. Our criteria was to stop the iterative process when the difference between two successive log-likelihoods is smaller than a pre-specified tolerance τ or when the number of non-zero coefficient estimates exceeds $N - 1$, and we set the defaults to $\epsilon = 0.001$ and $\tau = 0.00001$. Further, recall that for the linear regression the intercept γ_0 is commonly omitted because the response is centered. However, for count models it is inappropriate to center the outcome as that would result in negative counts. Therefore, the intercept must be included in the model without penalization, in addition to any covariates in the unpenalized subset. Once the iterative process has completed, the output includes a solution path for each coefficient. A 'final' model can then be selected from the resulting solution path based on predetermined desired criteria, such as the model attaining the minimum AIC, minimum BIC, or minimum cross-validated error.

Simulation studies

Simulation studies were performed to compare our negative binomial GMIFS model to existing penalized count response models including `glmpath` [68], `glmnet` [54], and `nnlasso` [64]. First, we randomly generated P predictor variables for observations $i = 1, \ldots, N$ from a standard normal distribution. Thereafter five of the P variables were selected to be associated with the discrete response and their coefficients were set to $\beta = \pm\log(\delta)$. We also assigned an intercept to take the value of $\gamma_0 = 0.5$ and we assigned a heterogeneity parameter α. We then calculated the mean response per observation as $\mu_i = \exp(\gamma_0 + \sum_{k=1}^{5} \beta_k x_{ik})$. The discrete response was then generated as $Y_i \sim$ Negative Binomial(μ_i, α). Once the discrete response was generated, we fit the following models: our negative binomial GMIFS model, a `glmpath` model with `family = poisson`, a `glmnet` model with `family = "poisson"`, and a `nnlasso` model with `family = "poisson"`. These methods maximize the log-likelihood with the additional penalty term λ placed on the regression coefficients, $\ell(\boldsymbol{\beta}; y, \alpha) - \lambda \sum_{p=1}^{P} |\beta_p|$. To ensure a fair comparison across the four modeling methods, we extracted the GMIFS model that attained the minimum AIC and minimum BIC, summed the absolute values of the estimated regression coefficients, then identified the step at which `glmpath`, `glmnet`, and `nnlasso` first attained a sum of the absolute value of the regression coefficients at the GMIFS level. This procedure was repeated $r = 200$ times. Simulations were performed using $N = 100$, $P = 500$, $\alpha = 0.3$ and 0.5, and $\delta = 1.5$ and $\delta = 1.75$. The four methods were compared with respect to the number of true predictors that had a non-zero coefficient estimate; the number of false predictors that had a non-zero coefficient estimate; and prediction error.

A larger simulation study consisting of ($P = 5,000$) correlated rather than independent features and $N = 50$ observations was also performed to mimic the structure of our application dataset. In the MoBa dataset, heterogeneity parameter estimate for the BIC-selected model was $\hat{\alpha}_{BIC} = 0.337$ and the six parameter estimates from the unpenalized model ranged from [-0.988, 1.122]. Therefore we set $\alpha = 0.35$, the intercept to $\gamma_0 = 0.5$, and conservatively set the parameter values of the true covariates to $\beta = \pm 0.693$, which corresponds to $\beta = \pm\log(2)$. We also estimated the correlations between genes included in the final model and all remaining genes in the dataset. This distribution is approximately normally distributed with a mean of 0.005 and a standard deviation of 0.28. Therefore, we developed a block diagonal correlation matrix with 40 features in each block and 125 blocks in total, to yield a 5000×5000 correlation matrix. First, the lower triangle of each 40×40 block of the correlation matrix was filled by generating random variates from a N(0,0.28) distribution. Second, the upper triangle was completed by enforcing the matrix to be symmetric. Third, the diagonal elements were taken to be 1. Thereafter, one feature from five different blocks was selected to represent the true covariates. The mean response per observation was taken to be $\mu_i = \exp(\gamma_0 + \sum_{k=1}^{5} \beta_k x_{ik})$ and the discrete response was then generated as $Y_i \sim$ Negative Binomial(μ_i, α). Once the discrete response was generated, we fit our negative binomial GMIFS model and the `glmnet` model with `family = "poisson"`. This procedure was repeated $r = 100$ times. Similar to [54], we omitted comparison to the `glmpath` algorithm because the algorithm does not scale well to the large size of this simulated dataset. Also, we omitted the comparison to `nnlasso` because, as demonstrated in the small simulation study, its non-negative constraint on the parameters results in poor performance when parameters can take both positive and negative values. The two methods were compared with respect to the number of true predictors that had a non-zero coefficient estimate; the number of false predictors that had a non-zero coefficient estimate; and prediction error. All simulations were performed on the Ohio Supercomputer Center cluster owens [69].

Norwegian Mother and Child Cohort Study

In the 1990s the Norwegian Mother and Child Cohort Study (MoBa) was designed collaboratively by researchers at the Medical Birth Registry of Norway (MBRN) and by researchers at the National Institute of Public Health [70]. Pregnant women who attended routine ultrasounds in Norway were recruited from 1999 to 2005 from 52 hospitals and maternity units. There was no exclusion criteria, and women who were pregnant more than once in the time period could participate multiple times. The pregnancy was defined as the unit of observation of the study. For 200 neonates, umbilical cord blood samples were collected immediately after birth. After quality control and other exclusions, 111 samples were hybridized to Agilent 4x44k human oligonucleotide microarrays to measure gene expression. Sample processing, image analysis, normalization, background correction, and filtering for the gene expression data are described in [71]. For an even smaller subset ($N = 29$), MN data were collected and scored using the procedure previously described [72]. Data were downloaded from Gene Expression Omnibus (GSE31836).

Results

Simulation studies

From the simulation studies we observed that the `nnlasso` method had poor performance, with the maximum number of false predictors included in the BIC-selected models extending to 292, 296, 298, and 291 when $\alpha = 0.3$ and $\beta = \pm\log(1.5)$ or $\pm\log(1.75)$ and for $\alpha = 0.5$ and

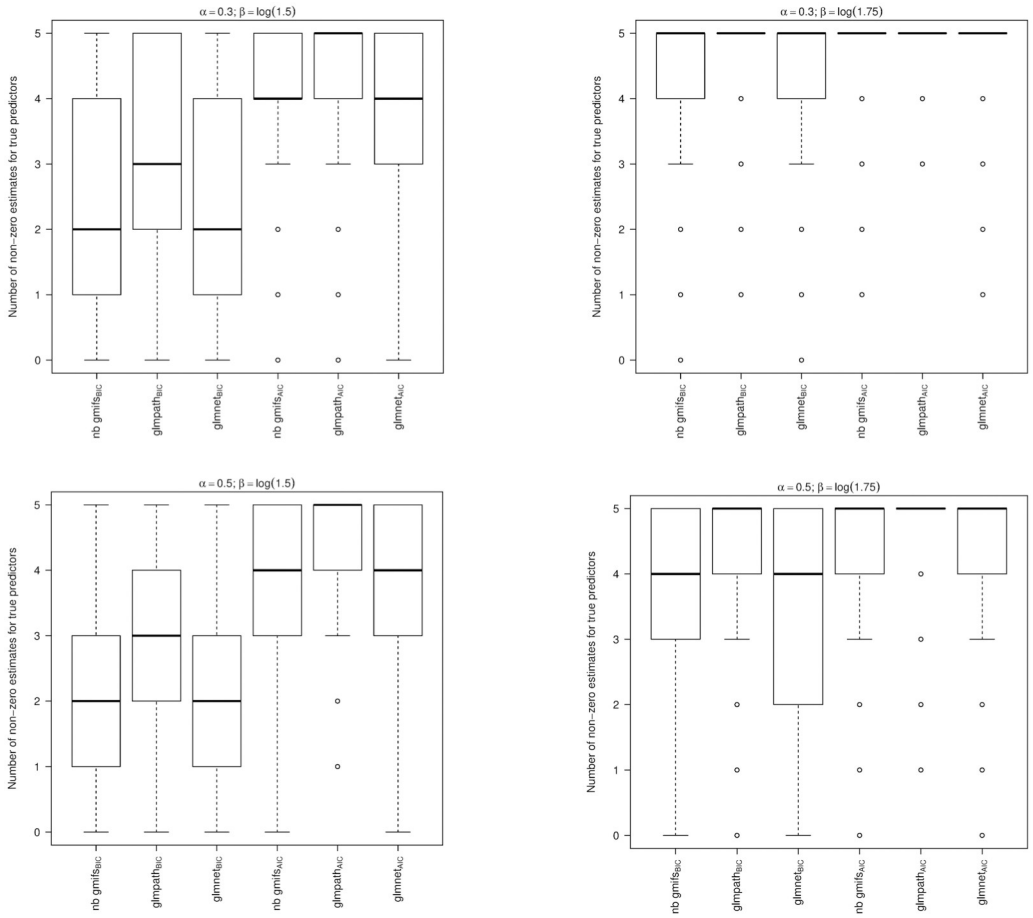

Fig 3. For each modeling method, number of true predictors that had a non-zero coefficient estimate (Oracle = 5). A: $\alpha = 0.3; \beta = \pm\log(1.5)$. B: $\alpha = 0.3; \beta = \pm\log(1.75)$. C: $\alpha = 0.5; \beta = \pm\log(1.5)$. D: $\alpha = 0.5; \beta = \pm\log(1.75)$.

https://doi.org/10.1371/journal.pone.0209923.g003

$\beta = \pm\log(1.5)$ or $\pm\log(1.75)$, respectively, which may be due to its non-negativity constraint on the parameter estimates. Therefore the `nnlasso` results were omitted from the boxplots because their inclusion obfuscated the results for the other methods. When examining the boxplots of the simulation results, we observed that our NB GMIFS method performed well with respect to identifying true predictors (Fig 3) while minimizing the number of false predictors included in the model (Fig 4), particularly as α and β increased. We also observed that `glmnet` performed comparable to our method, though our NB GMIFS method performed better than `glmnet` when identifying true predictors for AIC selected models when $\alpha = 0.3$ and $\beta = \pm\log(1.5)$ and for BIC selected models when $\alpha = 0.5$ and $\beta = \pm\log(1.75)$ (Fig 3). Additionally, our NB GMIFS method identified fewer false predictors, especially for the two $\beta = \pm\log(1.75)$ scenarios (Fig 4). While `glmpath` performed well at selecting the true predictors, it over-selected false predictors (Fig 4). Because our negative binomial GMIFS model had good performance with respect to prediction error (Fig 5) and minimized the number of false predictors included while doing well at selecting the true predictors, it is preferred for modeling an outcome that is over-dispersed.

When examining the results from the larger simulation that consisted of $P = 5,000$ correlated covariates for $N = 50$ subjects, we observed similar performance between our NB GMIFS method and `glmnet` with respect to identifying true predictors (Fig 6). However, our NB

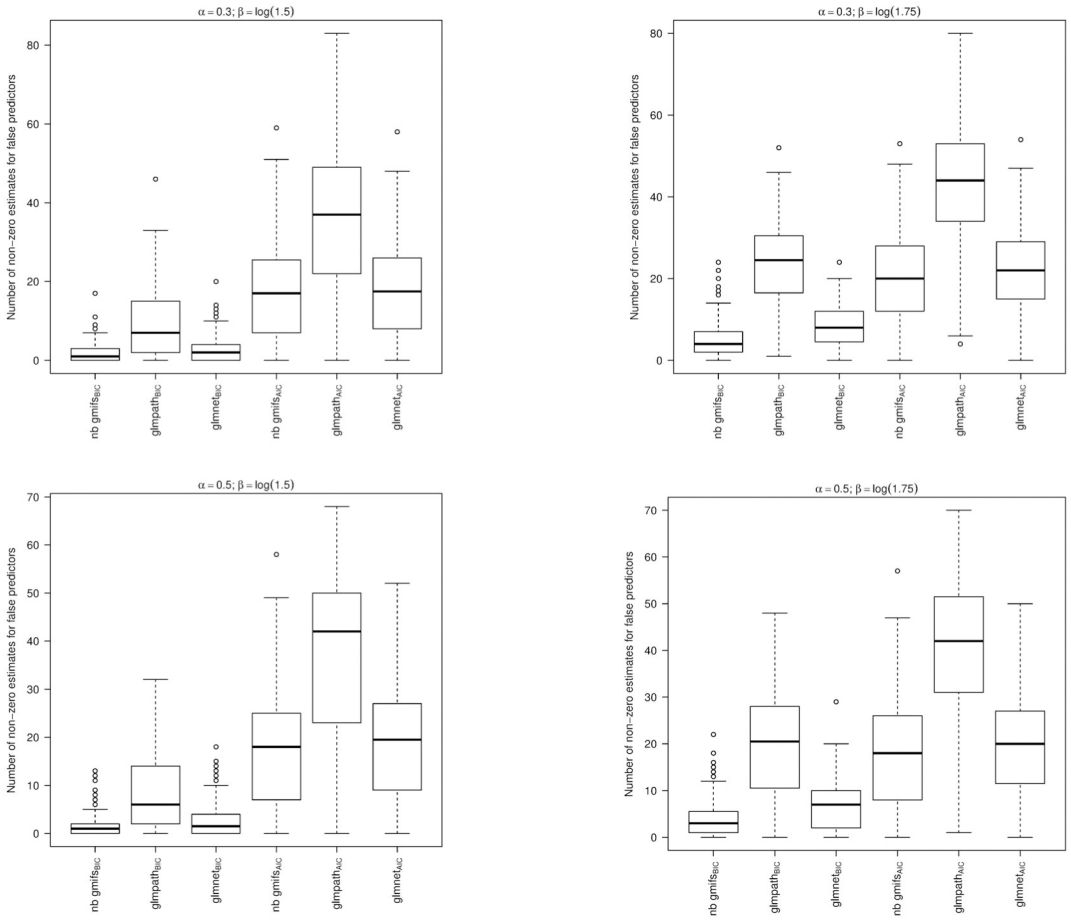

Fig 4. For each modeling method, number of false predictors that had a non-zero coefficient estimate (Oracle = 495). A: $\alpha = 0.3; \beta = \pm\log(1.5)$. B: $\alpha = 0.3; \beta = \pm\log(1.75)$. C: $\alpha = 0.5; \beta = \pm\log(1.5)$. D: $\alpha = 0.5; \beta = \pm\log(1.75)$.

https://doi.org/10.1371/journal.pone.0209923.g004

GMIFS method included fewer false predictors than `glmnet` (Wilcoxon signed rank test $P < 0.0001$ and $P < 0.0001$ for the AIC- and BIC-selected models, respectively) (Fig 6).

Norwegian Mother and Child Cohort Study

Before statistical analysis, a Boundary Likelihood Ratio test was performed to determine whether a Poisson or negative binomial model would be more appropriate given the MoBa data [38]. The alternative hypothesis of $\alpha \neq 0$ was tested against a null hypothesis of $\alpha = 0$. The chi-square test yielded $\chi^2_1 = 59.8$ which corresponds to a p-value of < 0.0001. Therefore, we rejected the null hypothesis that $\alpha = 0$ and concluded a negative binomial model is more appropriate given the data. When applying the score test to test the null hypothesis of no over-dispersion against the alternative hypothesis that over-dispersion is present, $P = 0.011$. When testing for over-dispersion using the Lagrange multiplier test, $P < 0.0001$; therefore, all tests indicated over-dispersion is present such that the NB model is preferred to the Poisson. Because the feature set is high-dimensional gene expression data, we used our GMIFS NB model to fit a multivariable model to predict MN frequency. We further filtered the gene expression dataset to include genes that had no missing values, leaving 8,497 genes for statistical modeling. Because the performance of `glmnet` and `glmpath` was somewhat comparable to our GMIFS method in the simulation studies, with `nnlasso` demonstrating poor

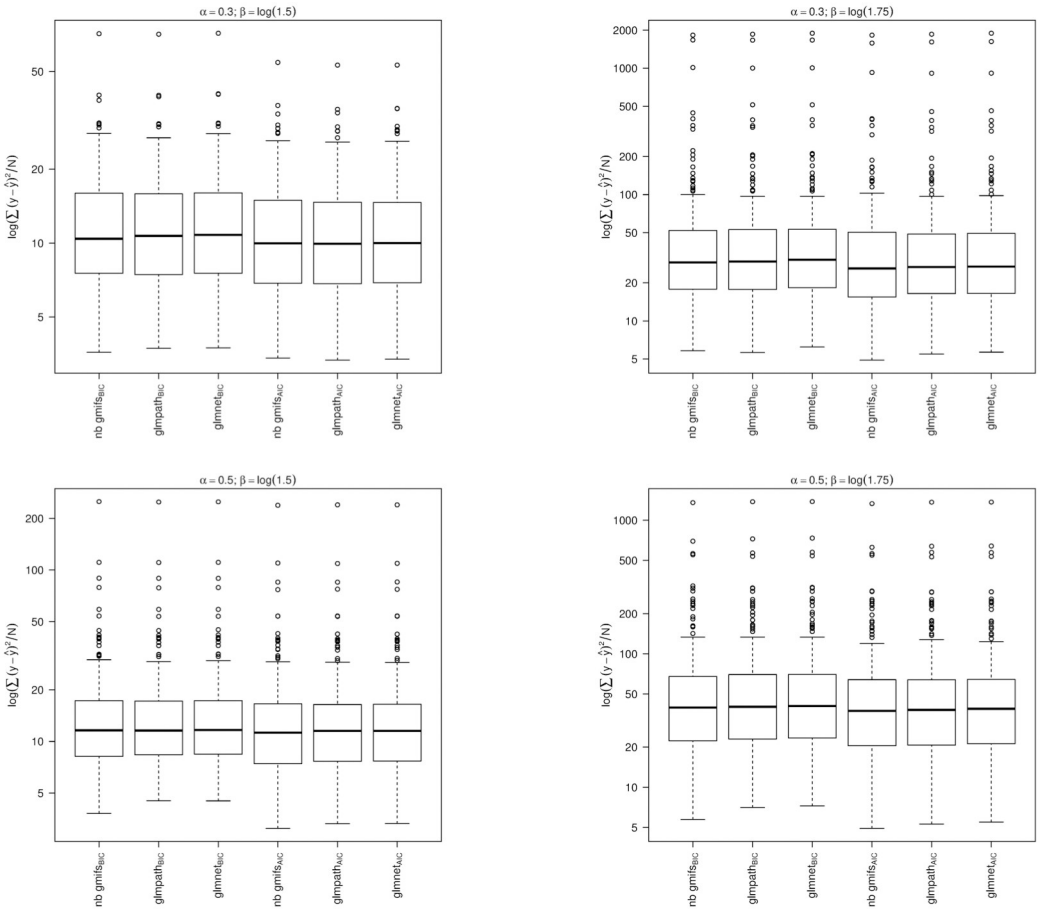

Fig 5. Prediction error for each modeling method and selection criteria. A: $\alpha = 0.3$; $\beta = \pm\log(1.5)$. B: $\alpha = 0.3$; $\beta = \pm\log(1.75)$. C: $\alpha = 0.5$; $\beta = \pm\log(1.5)$. D: $\alpha = 0.5$; $\beta = \pm\log(1.75)$.

https://doi.org/10.1371/journal.pone.0209923.g005

performance likely owing to its non-negativity constraint, we applied `glmnet` and `glmpath` to the MoBa data as comparative methods.

Though maternal age, gestational age, and maternal smoking status would ordinarily be of interest to include a unpenalized predictors, those data were not available so the only unpenalized predictor included in our model was neonate gender. The gene expression data were included in the model as penalized predictors. There were 13 genes with non-zero coefficient estimates in the AIC selected NB GMIFS model and six in the BIC selected NB GMIFS model (Table 1). The BIC attained a minimum at step 580 while the AIC attained its minimum at step 1102. The AIC selected `glmpath` Poisson model included 23 genes while the BIC selected `glmpath` Poisson model included 17 genes, so similar to our simulation studies, `glmpath` seems to overfit. Nine of the genes from the AIC selected NB GMIFS and `glmpath` Poisson models overlapped. Again, because `glmnet` does not include functions or relevant output for estimating AIC and BIC, we extracted the GMIFS models that attained the minimum AIC and minimum BIC, summed the absolute values of the estimated regression coefficients, then identified the step at which `glmnet` first attained a sum of the absolute value of the regression coefficients at the GMIFS level. The AIC-like `glmnet` Poisson model included nine genes while the BIC selected `glmnet` Poisson model included five genes; seven

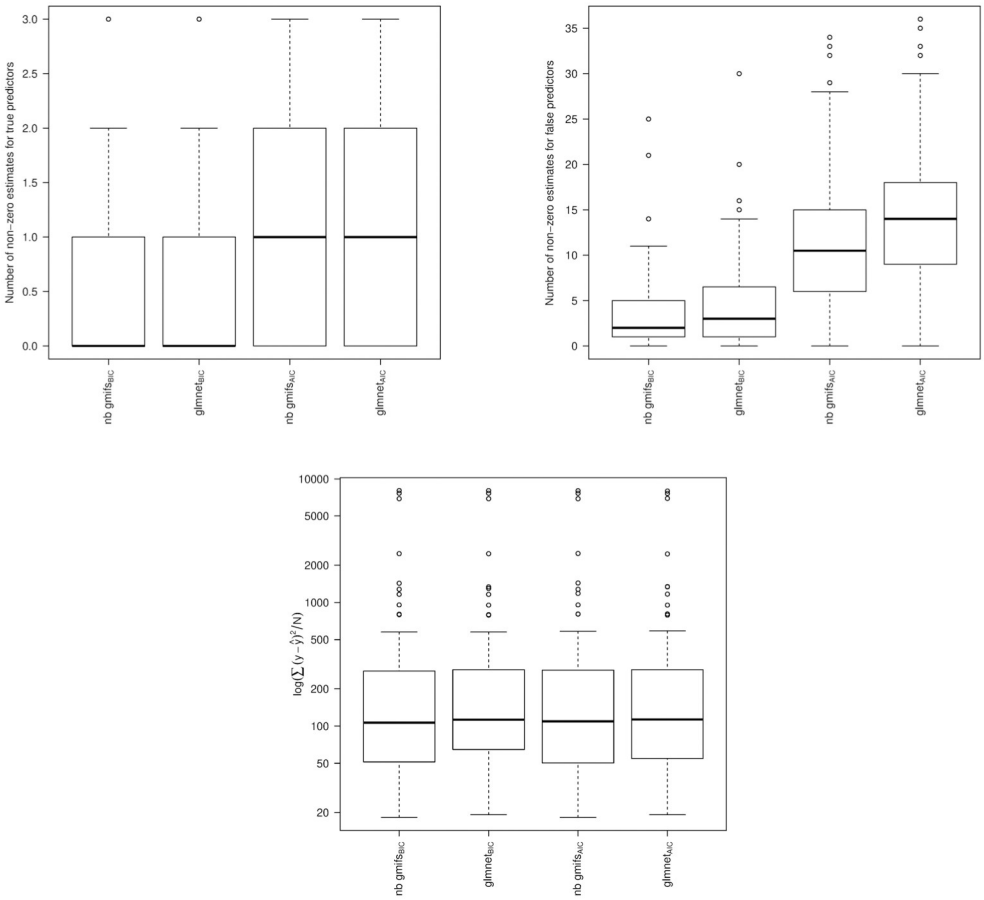

Fig 6. Results from large simulation study of $P = 5,000$ correlated predictors and $N = 50$ observations. A: Number of true predictors that had a non-zero coefficient estimate (Oracle = 5). B: Number of false predictors that had a non-zero coefficient estimate (Oracle = 4,995). C: Prediction error for each modeling method and selection criteria.

https://doi.org/10.1371/journal.pone.0209923.g006

of the genes from the AIC selected NB GMIFS and `glmnet` Poisson models overlapped. As suspected, the monotone solution paths for NB GMIFS are much smoother than those produced by `glmnet`, likely due to the correlated nature of the covariates (Fig 7). While the mean prediction error for the AIC- and BIC-selected models from our NB GMIFS algorithm were 2.21 and 3.23, respectively, the mean prediction errors were 3.86 and 3.30 for `glmnet`. As expected, the generalization error was larger as the N-fold cross-validation estimates of mean prediction error were 5.40 and 5.18 for the AIC- and BIC-selected NB GMIFS models, respectively, and 5.42 and 4.99 for the AIC and BIC `glmnet` models, respectively.

Interestingly, genes included in all models predicting MN frequency have been previously linked to relevant disease processes and cancer. For example, USP10 has been previously found to be involved in autophagy [73] and DNA damage response of cells [74]. CBX7 has been associated with thyroid [75] and endometrial cancer [76]. WHSC1, which is a synonym for NSD2, is involved in morphogenesis of anatomic structure [77] and associated with hematologic malignancies [77–79] and hepatocellular carcinoma [80]. KIAA0258 is a synonym RGP1. A mutation in RGP1 has been associated with adenocarcinoma of the large intestine [81]. C21ORF57 is a synonym for YBEY which according to COSM1031614, mutations in this gene in two TCGA samples have been associated with endometrioid carcinoma.

Table 1. Genes associated with MN frequency in the AIC and BIC selected NB GMIFS models.

Accession ID	Gene Symbol	Gene name	NB GMIFS AIC	NB GMIFS BIC	glmpath AIC	glmpath BIC	glmnet AIC	glmnet BIC
A_23_P100196	USP10	ubiquitin specific peptidase 10	X	X	X	X	X	X
A_23_P103824	THC2249577						X	
A_23_P133424	SKP1		X					
A_23_P138967	SDHD	succinate dehydrogenase complex	X		X			
A_23_P209394	CFLAR	CASP8 and FADD-like apoptosis regulator	X				X	
A_23_P42331	HMGA1	high mobility group AT-hook 1	X		X			
A_23_P79911	PSMF1	proteasome (prosome, macropain) inhibitor subunit 1 (PI31)					X	
A_24_P19410	CBX7	chromobox homolog 7	X	X	X	X	X	
A_24_P214858	TREML2	triggering receptor expressed on myeloid cells-like 2	X		X			
A_24_P2463	WHSC1	NSD2, Wolf-Hirschhorn syndrome candidate 1	X	X	X	X	X	X
A_24_P333019	RNF24	ring finger protein 24	X					
A_24_P397584	TBCC	tubulin folding cofactor C	X		X			
A_24_P398064	KIAA0258	RGP1 homolog, RAB6A GEF complex partner 1	X	X	X	X	X	X
A_32_P156549	C1ORF144		X	X			X	X
A_32_P18547	C21ORF57	chromosome 21 open reading frame 57	X	X	X	X	X	X

https://doi.org/10.1371/journal.pone.0209923.t001

Discussion

The simulation studies established that when the underlying data follow a negative binomial distribution (that is, the outcome is over-dispersed) the negative binomial GMIFS model outperforms penalized Poisson models with respect to including fewer false predictors. It also includes a substantial number of true predictors, particularly when the strength of association between the outcome and the predictor variable increases. Often it is of interest to account for the denominator, e.g., the number of binucleated cells scored, in the model, particularly if it varies by subject. This is usually accommodated by generalized linear modeling software by specifying the denominator to be an offset. While `glmpath` and `glmnet` include a parameter to the function call that allows for inclusion on an offset, both suffered from convergence issues when including an offset term. Also, the current implementation of `nnlasso` does not permit inclusion of an offset, which we consider to be a limitation of that package. Inclusion of

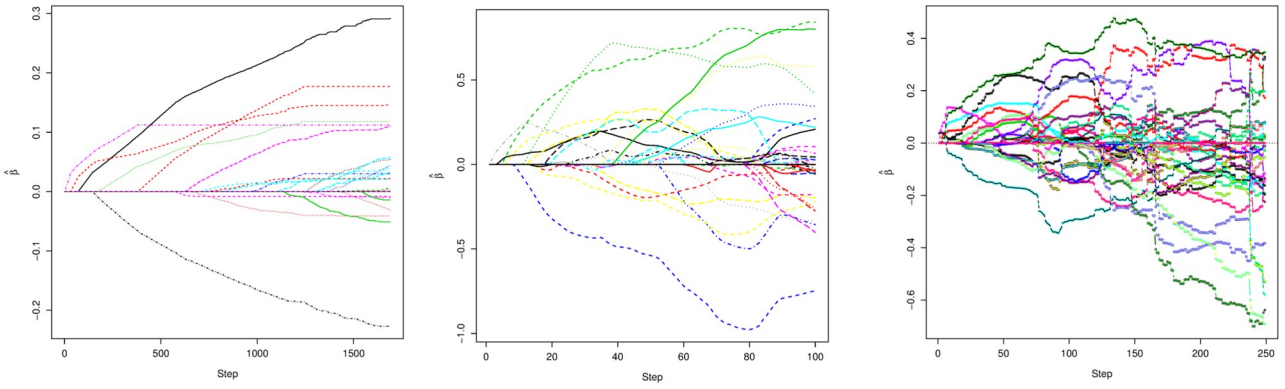

Fig 7. Coefficient paths for each modeling method for the MoBa dataset. A: Coefficient path from NB GMIFS. B: Coefficient path from `glmnet` where the x-axis indicates the sequence number of λ. C: Coefficient path from `glmpath`.

https://doi.org/10.1371/journal.pone.0209923.g007

an offset is not problematic for our GMIFS procedure given its incremental nature. Therefore our negative binomial GMIFS model offers several advantages.

Using various goodness-of-fit tests, it was determined that the micronuclei frequencies observed in the MoBa study more closely followed a negative binomial rather than a Poisson distribution. That finding is consistent with what has previously been reported [37]. When our NB GMIFS model was applied to the MoBa dataset, interesting genes that have previously been associated with cancer or relevant processes were identified, though genes that are truly associated with MN frequency in the MoBa study are unknown. One limitation is that when estimating Pearson's correlation, $\hat{\rho}$, between all remaining genes and the 13 genes included in the AIC-selected model, 316 were significantly correlated at an FDR< 0.05. Although the distribution of the correlation estimates appeared Gaussian and centered near zero ($\bar{\rho} = 0.005$ and $\hat{\sigma}_{\rho} = 0.28$), among the 316 significant genes, the absolute value of the estimated correlations ranged from 0.718 to 0.889. Therefore, it may be of scientific interest to explore the biological functions of these 316 genes because penalized methods tend to omit variables from the model if good proxies are already included.

Conclusion

Though the proposed NB GMIFS method was conceived of by considering high-throughput genomic applications, it is broadly applicable to a variety of health, social, and behavioral research fields, which commonly collect responses on an discrete scale. For example, our method is broadly applicable for modeling other discrete responses in high-dimensions, using the large number of predictors from an Electronic Medical Record system, for outcomes including but not limited to: length of hospital stay, number of drinks per day, and number of positive lymph nodes. An essential aspect for the dissemination of statistical methods is the development of software for the scientific user community. Therefore, our `countgmifs` package for the widely used `R` programming environment is available for download from the Comprehensive R Archive Network for others to model a discrete response, using either our Poisson of NB GMIFS method, when the covariate space is high-dimensional.

Author Contributions

Conceptualization: Kellie J. Archer.

Formal analysis: Kellie J. Archer.

Methodology: Rebecca R. Lehman, Kellie J. Archer.

Software: Rebecca R. Lehman, Kellie J. Archer.

Supervision: Kellie J. Archer.

Writing – original draft: Rebecca R. Lehman, Kellie J. Archer.

Writing – review & editing: Rebecca R. Lehman, Kellie J. Archer.

References

1. Tucker JD, Preston RJ. Chromosome aberrations, micronuclei, aneuploidy, sister chromatid exchanges, and cancer risk assessment. Mutat Res. 1996; 365: 147–159. https://doi.org/10.1016/S0165-1110(96)90018-4 PMID: 8898995
2. Fenech M, Morley AA. Measurement of micronuclei in lymphocytes. Mutat Res. 1985; 147: 29–36. https://doi.org/10.1016/0165-1161(85)90015-9 PMID: 3974610

3. Karahalil B, Karakaya AE, Burgaz S. The micronucleus assay in exfoliated buccal cells: Application to occupational exposure to polycyclic aromatic hydrocarbons. Mutat Res. 1999; 442: 29–35. https://doi.org/10.1016/S1383-5718(99)00055-8 PMID: 10366770
4. Yang HY, Feng R, Liu J, Wang HY, Wang YD. Increased frequency of micronuclei in binucleated lymphocytes among occupationally pesticide-exposed populations: A meta-analysis. Asian Pac J Cancer Prev. 2014; 15: 6955–6960. https://doi.org/10.7314/APJCP.2014.15.16.6955 PMID: 25169553
5. Warner ML, Moore LE, Smith MT, Kalman DA, Fanning E, Smith AH. Increased micronuclei in exfoliated bladder cells of individuals who chronically ingest arsenic-contaminated water in Nevada. Cancer Epidemiol Biomarkers Prev. 1994; 3: 583–590. PMID: 7827589
6. Holland N, Bolognesi C, Kirsch-Volders M, Bonassi S, Zeiger E, Knasmueller S, et al. The micronucleus assay in human buccal cells as a tool for biomonitoring DNA damage: The HUMN project perspective on current status, knowledge gaps. Mutat Res. 2008; 659: 93–108. https://doi.org/10.1016/j.mrrev.2008.03.007 PMID: 18514568
7. Bonassi S, Znaor A, Ceppi M, Lando C, Chang WP, Holland N, et al. An increased micronucleus frequency in peripheral blood lymphocytes predicts the risk of cancer in humans. Carcinogenesis 2007; 28(3): 625–631. https://doi.org/10.1093/carcin/bgl177 PMID: 16973674
8. Kirsch-Volders M, Decordier I, Elhajouji A, Plas G, Aardema MJ, Fenech M. In vitro genotoxicity testing using the micronucleus assay in cell lines, human lymphocytes, 3D human skin models. Mutagenesis 2011; 26(1): 177–184. https://doi.org/10.1093/mutage/geq068 PMID: 21164200
9. El-Zein RA, Fenech M, Lopez MS, Spitz MR, Etzel CJ. Cytokinesis-blocked micronucleus cytome assay biomarkers indentify lung cancer cases amongst smokers. Cancer Epidemiol Biomarkers Prev. 2008; 15(5): 1111–1119. https://doi.org/10.1158/1055-9965.EPI-07-2827
10. Varga D, Hoegel J, Maier C, Jainta S, Hoene M, Patino-Garcia B, et al. On the difference of micronucleus frequencies in peripheral blood lymphocytes between breast cancer patients, controls. Mutagenesis 2006; 21(5): 313–320. https://doi.org/10.1093/mutage/gel035 PMID: 16928695
11. Aristei C, Stracci F, Guerrieri P, Anselmo P, Armellini R, Rulli A, et al. Frequency of sister chromatid exchanges, micronuclei monitored over time in patients with early-stage breast cancer: Results of an observational study. Cancer Genet Cytogenet. 2009; 192: First.24–29. https://doi.org/10.1016/j.cancergencyto.2009.02.019 PMID: 19480933
12. Santos RA, Teixeira AC, Mayorano MB, Carrara HHA, Andrade JM, Takahashi CS. Basal levels of DNA damage detected by micronuclei, comet assays in untreated breast cancer patients, healthy women. Clin Exp Med. 2010; 10: 87–92. https://doi.org/10.1007/s10238-009-0079-4 PMID: 19902326
13. Flores-García A, Torres-Bugarín O, Velarde-Félix JS, Rangel-Villalobos H, Zepeda-Carrillo EA, Rodríguez-Trejo A, et al. Micronuclei, other nuclear anomalies in exfoliated buccal mucosa cells of Mexican women with breast cancer. J BUON 2014; 19(4): 895–899. PMID: 25536592
14. Celik DA, Kosar PA, Özcelik N, Eroglu E. Cytogenetic finding of breast cancer cases, in their first-degree relatives. J Breast Cancer 2013; 16(3): 285–290. https://doi.org/10.4048/jbc.2013.16.3.285 PMID: 24155757
15. Lee SJ, Yum YN, Kim SC, Kim Y, Lim J, Lee WJ, et al. Distinguishing between genotoxic, non-genotoxic hepatogarcinogens by gene expression profiling, bioinformatic pathway analysis. Sci Rep. 2013; 3: 2783. PMID: 24089152
16. Hond ED, Govarts E, Willems H, Smolders R, Casteleyn L, Kolossa Gehring M, et al. First steps toward harmonized human biomonitoring in Europe: Demonstration project to perform human biomonitoring on a European scale. Environ Health Perspect. 2015; 123(3): 255–263. https://doi.org/10.1289/ehp.1408616
17. Surowy H, Rinckleb A, Luedeke M, Stuber M, Wecker A, Varga D, et al. Heritability of baseline, induced micronucleus frequencies. Mutagenesis 2011; 26(1): 111–117. https://doi.org/10.1093/mutage/geq059 PMID: 21164191
18. Sagari SG, Babannavar R, Lohra A, Kodgi A, Bapure S, Rao Y, et al. Micronuclei frequencies, nuclear abnormalities in oral exfoliated cells of nuclear power plant workers. J Clin Diagn Res. 2014; 8(12): ZC15–ZC17. https://doi.org/10.7860/JCDR/2014/9059.5240 PMID: 25654022
19. Moretti M, Grollino MG, Pavanello S, Bonfiglioli R, Villarini M, Appolloni M, et al. Micronuclei, chromosome aberrations in subjects occupationally exposed to antineoplastic drugs: A multicentric approach. Int Arch Occup Environ Health 2015; 88(6): 683–695. https://doi.org/10.1007/s00420-014-0993-y PMID: 25362515
20. Minozzo R, Deimling LI, Gigante LP, Santos-Mello R. Micronuclei in peripheral blood lymphocytes of workers exposed to lead. Mutat Res. 2004; 565: 53–60. https://doi.org/10.1016/j.mrgentox.2004.09.003 PMID: 15576239

21. Gandhi G, Kaur W. Micronucleus frequencies in exfoliated urothelial cells among individuals residing near a waste water drain, using underground water resources. Toxicol Mech Methods 2005; 15: 219–225. https://doi.org/10.1080/15376520590945649 PMID: 20021086

22. Banerjee M, Banerjee N, Bhattacharjee P, Mondal D, Lythgoe PR, Martínez M, et al. High arsenic in rice is associated with elevated genotoxic effects in humans. Sci Rep. 2013; 3(2195): 1–8.

23. Merlo DF, Ceppi M, Stagi E, Bocchini V, Sram RJ, Rossner P. Baseline chromosome aberrations in children. Toxicol Lett. 2007; 172: 60–67. https://doi.org/10.1016/j.toxlet.2007.05.016 PMID: 17604577

24. Pedersen M, Vinzents P, Petersen JH, Kleinjans JCS, Plas G, Kirsch-Volders M, et al. Cytogenetic effects in children, mothers exposed to air pollution assessed by the frequency of micronuclei, fluorescence in situ hybridization (FISH): A family pilot study in the Czech Republic. Mutat Res. 2006; 608: 112–120. https://doi.org/10.1016/j.mrgentox.2006.02.013 PMID: 16829164

25. Battershill JM, Burnett K, Bull S. Factors affecting the incidence of genotoxicity biomarkers in peripheral blood lymphocytes: Impact on design of biomonitoring studies. Mutagenesis 2008; 23(6): 423–427. https://doi.org/10.1093/mutage/gen040 PMID: 18678752

26. Cassel APR, Barcellos RB, de Silva CMD, de Matos Almeida SE, Rossetti MLR. Association between human papillomavirus (HPV) DNA, micronuclei in normal cervical cytology. Genet Mol Biol. 2014; 37(2): 360–363. https://doi.org/10.1590/S1415-47572014005000010 PMID: 25071400

27. Gutierrez S, Carbonell E, Galofre P, Creus A, Marcos R. Cytogenetic damage after 131-iodine treatment for hyperthyroidism, thyroid cancer. Eur J Nucl Med. 1999; 26(12): 1589–1596.

28. Samanta S, Pey P, Nijhawan R. The role of micronucleus scoring in fine needle aspirates of ductal carcinoma of the breast. Cytopathology 2011; 22: 111–114. https://doi.org/10.1111/j.1365-2303.2010.00773.x PMID: 20553316

29. Hemalatha A, Suresh TN, HarendraKumar ML. Micronuclei in breast aspirates. Is scoring them helpful? J Cancer Res Ther. 2014; 10(2): 309–311. https://doi.org/10.4103/0973-1482.136588 PMID: 25022383

30. Cardinale F, Bruzzi P, Bolognesi C. Role of micronucleus test in predicting breast cancer susceptibility: A systematic review, meta-analysis. Br J Cancer 2012; 106: 780–790. PMID: 22187037

31. Fischer WH, Keiwan A, Schmitt E, Stopper H. Increased formation of micronuclei after hormonal stimulation of cell proliferation in human breast cancer cells. Mutagenesis 2001; 16(3): 209–212. https://doi.org/10.1093/mutage/16.3.209 PMID: 11320145

32. Fenech M, Chang WP, Kirsch-Volders M, Holland N, Bonassi S, Zeiger E. HUMN project: Detailed description of the scoring criteria for the cytokinesis-block micronucleus assay using isolated human lymphocyte cultures. Mutat Res. 2003; 534: 65–75. https://doi.org/10.1016/S1383-5718(02)00249-8 PMID: 12504755

33. Fenech M, Kirsch-Volders M, Natarajan AT, Surralles J, Crott JW, Parry J, et al. Molecular mechanisms of micronucleus, nucleoplasmic bridge, nuclear bud formation in mammalian, human cells. Mutagenesis 2011; 26(1): 125–132. https://doi.org/10.1093/mutage/geq052 PMID: 21164193

34. Fenech M. Cytokinesis-block micronucleus cytome assay. Nat Protoc. 2007; 2(5): 1084–1104. https://doi.org/10.1038/nprot.2007.77 PMID: 17546000

35. Tolbert PE, Shy CM, Allen JW. Micronuclei, other nuclear anomalies in buccal smears: methods development. Mutat Res. 1992; 271: 69–77. https://doi.org/10.1016/0165-1161(92)90033-I PMID: 1371831

36. Thomas P, Holland N, Bolognesi C, Kirsch-Volders M, Bonassi S, Zeiger E, et al. Buccal micronucleas cytome assay. Nat Protoc. 2009; 4(6): 825–837. PMID: 19444240

37. Ceppi M, Biasotti B, Fenech M, Bonassi S. Human population studies with the exfoliated buccal micronucleus assay: Statistical, epidemiological issues. Mutat Res Rev Mutat Res. 2010; 705: 11–19. https://doi.org/10.1016/j.mrrev.2009.11.001

38. Hilbe JM. Negative Binomial Regression. Cambridge, UK: Cambridge University Press; 2011.

39. Ramirez A, Saldanha PH. Micronucleus investigation of alcoholic patients with oral carcinomas. Genet Mol Res. 2002; 1(3): 246–260. PMID: 14963832

40. Zeller J, Neuss S, Mueller JU, Kühner S, Holzmann K, Högel J, et al. Assessment of genotoxic effects and changes in gene expression in humans exposed to formaldehyde by inhalation under controlled conditions. Mutagenesis 2011; 26(4): 555–561. https://doi.org/10.1093/mutage/ger016 PMID: 21460374

41. Ghandhi SA, Ponnaiya B, Panigrahi SK, Hopkins KM, Cui Q, Hei TK, et al. RAD9 deficiency enhances radiation induced bystander DNA damage and transcriptomal response. Radiat Oncol. 2014; 9: 206. https://doi.org/10.1186/1748-717X-9-206 PMID: 25234738

42. Tibshirani R. Regression shrinkage, selection via the Lasso. J R Stat Soc Series B Stat Methodol 1996; 58: 267–288.

43. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997; 16: 385–395. https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4%3C385::AID-SIM380%3E3.0.CO;2-3 PMID: 9044528

44. Wu B. Differential gene expression detection using penalized linear regression models: the improved SAM statistics. Bioinformatics 2005; 21: 1565–1571. https://doi.org/10.1093/bioinformatics/bti217 PMID: 15598833

45. Wu B. Differential gene expression detection, sample classification using penalized linear regression models. Bioinformatics 2006; 22: 472–476. https://doi.org/10.1093/bioinformatics/bti827 PMID: 16352654

46. Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. Biostatistics 2004; 5: 427–443. https://doi.org/10.1093/biostatistics/5.3.427 PMID: 15208204

47. Schimek MG. Penalized binary regression for gene expression profiling. Methods Inf Med 2004; 43: 439–444. https://doi.org/10.1055/s-0038-1633894 PMID: 15702197

48. Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data. Bioinformatics 2007; 23: 1768–1774. https://doi.org/10.1093/bioinformatics/btm232 PMID: 17485430

49. Gui J, Li H. Penalized Cox regression analysis in the high-dimensional, low-sample size settings, with applications to microarray gene expression data. Bioinformatics 2005; 21: 3001–3008. https://doi.org/10.1093/bioinformatics/bti422 PMID: 15814556

50. Li H, Gui J. Partial Cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics 2004; 20 Suppl 1: i208–15. https://doi.org/10.1093/bioinformatics/bth900 PMID: 15262801

51. Shedden K, Taylor JMG, Enkemann SA, Tsao MS, Yeatman TJ, Gerald WL, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med. 2008; 14: 822–827. PMID: 18641660

52. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat. 2004; 32(2): 407–451. https://doi.org/10.1214/009053604000000067

53. Park MY, Hastie T. L_1-regularization path algorithm for generalized linear models. J R Stat Soc Series B Stat Methodol 2007; 69(Part 4): 659–677. https://doi.org/10.1111/j.1467-9868.2007.00607.x

54. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33(1): 1–22. https://doi.org/10.18637/jss.v033.i01 PMID: 20808728

55. Hastie T, Taylor J, Tibshirani R, Walther G. Forward stagewise regression, the monotone lasso. Electron J Stat. 2007; 1: 1–29. https://doi.org/10.1214/07-EJS004

56. Archer KJ, Williams AAA. L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. Stat Med. 2012; 31: 1464–74. https://doi.org/10.1002/sim.4484 PMID: 22359384

57. Archer KJ, Hou J, Williams AAA. Classifying normal, nevus, and malignant melanoma skin samples using penalized ordinal regression. New Frontiers of Multidisciplinary Research in STEAM-H (Science, Technology, Engineering, Agriculture, Mathematics, Health) Toni Bourama (editor) 2014; Springer Proceedings in Mathematics & Statistics:111–133.

58. Archer KJ, Hou J, Zhou Q, Ferber K, Layne JG, Gentry AE. ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. Cancer Inform. 2014; 13: 187–195. https://doi.org/10.4137/CIN.S20806 PMID: 25574124

59. Hou J, Archer KJ. Regularization method for predicting an ordinal response using longitudinal high-dimensional genomic data. Stat Appl Genet Mol Biol. 2015; 14(1): 93–111. https://doi.org/10.1515/sagmb-2014-0004 PMID: 25720102

60. Ferber K, Archer KJ. Modeling discrete survival time using genomic feature data. Cancer Inform. 2015; 14(Suppl2): 37–43. https://doi.org/10.4137/CIN.S17275 PMID: 25861216

61. Elswick Gentry A, Jackson-Cook C, Lyon D, Archer KJ. Penalized ordinal regression methods for predicting stage of cancer in high-dimensional covariate spaces. Cancer Inform. 2015; 14(Suppl2): 201–208.

62. Makowski M, Archer KJ. Generalized monotone incremental forward stagewise method for modeling count data: Application predicting micronuclei frequency. Cancer Inform. 2015; 14(Suppl2): 97–105. https://doi.org/10.4137/CIN.S17278 PMID: 25983544

63. Payne EH, Hardin JW, Egede IE, Ramakrishnan V, Selassie A, Gebregziabher M. Approaches for dealing with various sources of overdispersion in modeling count data: Scale adjustment versus modeling. Stat Methods Med Res. 2017; 26(4): 1802–1823. https://doi.org/10.1177/0962280215588569 PMID: 26031359

64. Mandal BN, Ma J. l_1 regularized multiplicative iterative path algorithm for non-negative generalized linear models Comput Stat Data Anal. 2016; 101: 289–299.

65. Schelldorfer J, Buhlmann P, van de Geer S. Estimation for high-dimensional linear mixed-effects models using l1-penalization. Scandinavian Journal of Statistics 2011; 38(2): 197–214. https://doi.org/10.1111/j.1467-9469.2011.00740.x

66. Schelldorfer J, Meier L, Buhlmann P. GLMMLasso: An algorithm for high-dimensional generalized linear mixed models using L1-penalization. J Comput Graph Stat. 2014; 23(2): 460–477. https://doi.org/10.1080/10618600.2013.773239

67. Tutz G, Groll A. Variable selection for generalized linear mixed models by L1-penalized estimation. Stat Comput. 2014; 24(2): 137–154. https://doi.org/10.1007/s11222-012-9359-z

68. Park MY, Hastie T. glmpath: L1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model. R package version 0.98, Available from: https://CRAN.R-project.org/package=glmpath, 2018.

69. Ohio Supercomputer Center Ohio Supercomputer Center. Columbus, OH: 1987. Available from: http://osc.edu/ark:/19495/f5s1ph73

70. Magnus P, Irgens LM, Haug K, Nystad W, Skjaerven R, Stoltenberg C, and The Moba Study Group. Cohort profile: The Norwegian Mother and Child Cohort Study (MoBa). Int J Epidemiol. 2006; 35(5): 1146–1150. https://doi.org/10.1093/ije/dyl170 PMID: 16926217

71. Hochstenbach K, van Leeuwen DM, Gmuender H, Gottschalk RW, Lovik M, Granum B, et al. Global gene expression analysis in cord blood reveals gender specific differences in response to carcinogenic exposure in utero. Cancer Epidemiol Biomarkers Prev. 2012; 21(10): 1756–1767. https://doi.org/10.1158/1055-9965.EPI-12-0304 PMID: 22879202

72. Decordier I, Papine A, Plas G, Roesems S, Vande Loock K, Moreno-Palomo J, et al. Automated image analysis of cytokinesis-block micronuclei: an adapted protocol and validated scoring procedure for biomonitoring. Mutagenesis 2009; 24(1): 85–93. https://doi.org/10.1093/mutage/gen057 PMID: 18854579

73. Liu J, Xia H, Kim M, Xu L, Li Y, Zhang L, et al. Beclin1 controls the levels of p53 by regulating the deubiquitination activity of USP10 and USP13. Cell 2011; 147(1): 223–234 https://doi.org/10.1016/j.cell.2011.08.037 PMID: 21962518

74. Wang W, Huang X, Xin HB, Fu M, Xue A, Wu ZH. TRAF Family Member-associated NF-κB Activator (TANK) Inhibits Genotoxic Nuclear Factor κB Activation by Facilitating Deubiquitinase USP10-dependent Deubiquitination of TRAF6 Ligase. J Biol Chem 2015; 290(21): 13372–85. https://doi.org/10.1074/jbc.M115.643767 PMID: 25861989

75. Federico A, Pallante P, Bianco M, Ferraro A, Esposito F, Monti M, et al. Chromobox protein homologue 7 protein, with decreased expression in human carcinomas, positively regulates E-cadherin expression by interacting with the histone deacetylase 2 protein. Cancer Res. 2009; 69(17): 7079–7087. https://doi.org/10.1158/0008-5472.CAN-09-1542 PMID: 19706751

76. Boren T, Xiong Y, Hakam A, Wenham R, Apte S, Wei Z, et al. MicroRNAs and their target messenger RNAs associated with endometrial carcinogenesis. Gynecol Oncol. 2008; 110(2): 206–15. https://doi.org/10.1016/j.ygyno.2008.03.023 PMID: 18499237

77. Stec I, Wright TJ, van Ommen GJ, de Boer PA, van Haeringen A, Moorman AF, et al. WHSC1, a 90 kb SET domain-containing gene, expressed in early development and homologous to a Drosophila dysmorphy gene maps in the Wolf-Hirschhorn syndrome critical region and is fused to IgH in t(4;14) multiple myeloma. Hum Mol Genet. 1998; 7(7): 1071–82. https://doi.org/10.1093/hmg/7.7.1071 PMID: 9618163

78. Kim JY, Kee HJ, Choe NW, Kim SM, Eom GH, Baek HJ, et al. Multiple-myeloma-related WHSC1/MMSET isoform RE-IIBP is a histone methyltransferase with transcriptional repression activity. Mol Cell Biol. 2008; 28(6): 2023–2034. https://doi.org/10.1128/MCB.02130-07 PMID: 18172012

79. He J, Abdel-Wahab O, Nahas MK, Wang K, Rampal RK, Intlekofer AM, et al. Integrated genomic DNA/RNA profiling of hematologic malignancies in the clinical setting. Blood 2016; 127(24): 3004–2014. https://doi.org/10.1182/blood-2015-08-664649 PMID: 26966091

80. Stefanska B, Huang J, Bhattacharyya B, Suderman M, Hallett M, et al. Definition of the landscape of promoter DNA hypomethylation in liver cancer. Cancer Res. 2011; 71(17): 5891–903. https://doi.org/10.1158/0008-5472.CAN-10-3823 PMID: 21747116

81. Giannakis M, Mu XJ, Shukla SA, Qian ZR, Cohen O, Nishihara R, et al. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. Cell Rep. 2016; 15(4): 857–865. https://doi.org/10.1016/j.celrep.2016.03.075 PMID: 27149842