2016

Evaluation of statistical methods, modeling, and multiple testing in RNA-seq studies

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

EVALUATION OF STATISTICAL METHODS, MODELING, AND MULTIPLE TESTING IN RNA-SEQ STUDIES

by

SEUNG HOAN CHOI

B.S., The State University of New York at Stony brook, 2008
M.A., Boston University, 2011

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2016

Approved by

First Reader ______________________________
Anita L Destefano, Ph.D.
Professor of Biostatistics

Second Reader ______________________________
Josée Dupuis, Ph.D.
Professor of Biostatistics

Third Reader ______________________________
Kathryn L Lunetta, Ph.D.
Professor of Biostatistics

To my beloved late grandfather

Hae Seul Choi (1922 – 2015)

ACKNOWLEDGMENTS

I would like to thank all my committee members for all their support toward this degree. My sincerest gratitude goes to my primary advisor, Dr. Anita L Destefano, for her all guidance and support throughout my time at Boston University. She always helped me think toward the right direction, obtain insights to see the whole picture, and concentrate on details where many may miss. This dissertation could not be finished without her expertise in statistical genetics and genomics and her insights.

I am very grateful to Dr. Josée Dupuis and Dr. Kathryn L Lunetta who shared their statistical expertise and supported my research. I would also like to thank Dr. Honghuang Lin and Dr. Richard H Myers for sharing ideas and offering advice and comments on my dissertation.

I also like to thank Dr. Sudha Seshadri, with whom I begin my research, for her support and guidance. Her enthusiasm toward research always motivates myself. With her support, I was enabled to participate in multiple large consortia and learn how to collaborate with other people.

I would like to thank my parents Geun Sik Choi and Hye Sook Kim and my sister Jina Choi for their unconditional love and support. Lastly, I would like to thank my

wife Kyung Hee Baek, to whom I dedicate my dissertation, for her love, support, and encouragement.

EVALUATION OF STATISTICAL METHODS, MODELING, AND MULTIPLE TESTING IN RNA-SEQ STUDIES

SEUNG HOAN CHOI

Boston University Graduate School of Arts and Sciences, 2016

Major Professor: Anita L Destefano, Professor of Biostatistics

ABSTRACT

Recent Next Generation Sequencing methods provide a count of RNA molecules in the form of short reads, yielding discrete, often highly non-normally distributed gene expression measurements. Due to this feature of RNA sequencing (RNA-Seq) data, appropriate statistical inference methods are required. Although Negative Binomial (NB) regression has been generally accepted in the analysis of RNA-Seq data, its appropriateness in the application to genetic studies has not been exhaustively evaluated. Additionally, adjusting for covariates that have an unknown relationship with expression of a gene has not been extensively evaluated in RNA-Seq studies using the NB framework. Finally, the dependent structures in RNA-Seq data may violate the assumptions of some multiple testing correction methods. In this dissertation, we suggest an alternative regression method, evaluate the effect of covariates, and compare various multiple testing correction methods. We conduct simulation studies and apply these methods to a real data set. First, we suggest Firth's logistic regression for detecting differentially expressed genes in RNA-Seq data. We also recommend the data adaptive method that estimates a recalibrated distribution of test statistics. Firth’ logistic regression exhibits an appropriately controlled Type-I error rate using the data adaptive method and shows comparable power to NB regression in simulation studies. Next, we evaluate the effect of disease-associated covariates where the relationship between the covariate and gene expression is unknown. Although the power of NB and Firth’s logistic regression is decreased as disease-associated covariates are added in a model, Type-I error rates are well controlled in Firth’ logistic regression if the relationship between a covariate and disease is not strong. Finally, we compare multiple testing correction methods that control family-wise error rates and impose false discovery rates. The evaluation reveals that an understanding of study designs, RNA-Seq data, and the consequences of applying specific regression and multiple testing correction methods are very important factors to control family-wise error rates or false discovery rates. We believe our statistical investigations will enrich gene expression studies and influence related statistical methods.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DE	Differential Expression
FDR	False Discovery Rate
FWER	Familywise Error Rate
NB	Negative Binomial
NCP	Non Confounding Predictive
NGS	Next Generation Sequencing
RNA	Ribonucleic Acid
RNA-Seq	RNA sequencing

Chapter 1 Introduction

1.1 Gene expression studies

Gene expression studies have played important roles to understand phenotypic variation including how tissues vary in gene expression and how these variations are related to biologic function(Ramsköld et al. 2009). Current next generation sequencing (NGS) genome-wide gene expression measurement methods simultaneously quantify tens of thousands of unique Ribonucleic Acid (RNA) molecules extracted from biological samples. These RNA sequencing (RNA-Seq) methods produce data that can be transformed into numerical values that are proportional to the abundance of RNA molecules of interest, including protein-coding, processed transcript, pseudo-genes, miRNAs, tRNAs, rRNAs, snRNAs, snoRNAs, and scRNAs(Tarazona et al. 2011), and represent the amount of expression of those molecules. A common task in the analysis of RNA-Seq data is to evaluate the statistical differences of the mean expression of genes between sets of samples from two different conditions, e.g. control versus diseased patients. Identifying differentially expressed genes is the first important step to understanding the molecular mechanism of the differentially expressed genes and developing novel therapies for related diseases.

Microarray technology has been widely used to measure gene expression in the past decades. Microarray technology quantifies the fluorescence of specific RNA molecules and, after processing and normalization, expression values are continuous and typically approximated by a normal distribution. Due to these characteristics of microarray data, well-understood methods like two sample t-tests and linear regression are often utilized to identify the association between expression level and disease status. In contrast, NGS methods provide a count of RNA molecules in the form of short reads, which are discrete measurements that do not follow a normal distribution. Consequently, statistical methods that assume normality are inappropriate for the analysis of these count data, and therefore the development of appropriate statistical methods is necessary.

Because the total number of reads of each sample will likely be different, a normalization step is required before analyzing the association between genes and a condition. Anders et al. proposed a normalization method that divides each count by the geometric mean count of the corresponding gene and takes the medians of these scaled counts within each library(Anders and Huber 2010). Robinson et al. developed the Trimmed Mean of M Values (TMM) method that computes each normalization factor from the trimmed mean of the gene-wise log fold-changes of the current library to a reference library(M. D. Robinson and Oshlack 2010). Mortazavi et al. suggested the standard reads per kilobase of transcript per million mapped reads (Mortazavi et al. 2008). An inappropriate normalization method may result in a biased differential expression (DE) inference(Bullard et al. 2010). Dillies et al. comprehensively evaluated normalization methods and stated that TMM and Anders et al’s methods provide similar and reasonable results in their evaluating metrics(Dillies et al. 2013). Appropriately normalized data allow us to perform unbiased differential expression inferences.

1.2 Negative Binomial regression

Poisson models are a popular approach to analyze count data observed from experiments or epidemics. Poisson models assume that the data follow a Poisson distribution, where the mean and variance are the same. When the variance is significantly larger than the mean, alternative models are required to analyze the over-dispersed count data. A common alternative approach is the Negative Binomial (NB) model, also known as the gamma-Poisson model.

This approach fits a NB generalized linear model (McCullagh and Nelder 1989) to the data with estimated or fixed value of a dispersion parameter. Let Y be the response variable and x be an explanatory variable. The marginal distribution of Y, and negative binomial likelihood are

$$Y\sim \mathrm{NB}(\mu(x),\phi), where\ \mu \geq 0\ and\ \phi \geq 0\ \text{ such that}$$

$$\Pr(Y = y|x) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)}\left(\frac{1}{1+\phi^{-1}\mu(x)}\right)^{\phi^{-1}}\left(\frac{\phi^{-1}\mu(x)}{1+\phi^{-1}\mu(x)}\right)^{y}, y = 0, 1, 2 \ldots ,$$

$$\mathrm{E}[Y|x] = \mu, and\ \mathrm{VAR}[Y|x] = \mu + \mu^2\phi\ .$$

When ϕ is close to zero, the distribution of Y becomes a Poisson distribution. Let $Y_i \sim \mathrm{NB}(\mu(x_i),\phi), i = 1, \ldots, n$ be independent, where $\mu(x_i) = \exp(x_i\beta)$ and x_i is the $p\times 1$ explanatory vector. The likelihood function is proportional to

$$L(\beta,\phi) = \prod_{i=1}^{n} \frac{\Gamma\left(y_i+\phi^{-1}\right)}{\Gamma(\phi^{-1})\Gamma(y_i+1)} \left(\frac{1}{1+\phi^{-1}\mu(x_i)}\right)^{\phi^{-1}} \left(\frac{\phi^{-1}\mu(x_i)}{1+\phi^{-1}\mu(x_i)}\right)^{y_i},$$

and the log $L(\beta,\phi)$ is

$$l(\beta,\phi) = \sum_{i=1}^{n} \left(\sum_{j=0}^{y_i-1} \log(1+\phi j) + y_i \log\big(\mu(x_i)\big) - (y_i - \phi^{-1})\log\left(1+\phi\mu(x_i)\right)\right).$$

The obtained $(\hat{\beta}_{ML}, \hat{\phi}_{ML})$ maximize $l(\beta,\phi)$ through scores and information iterations(McCullagh and Nelder 1989). However, in general, a variance parameter from maximum likelihood estimators is underestimated (M. D. Robinson and Smyth 2007), hence alternative methods are suggested for the estimation of ϕ.

The pseudo-likelihood model (Breslow 1984) estimates the variance parameter using a distribution free goodness-of-fit statistic by solving the moment function

$$\sum_{i=1}^{n} \frac{\left(y_i-\hat{\mu}_{ML,i}\right)^2}{\hat{\mu}_i(1+\hat{\phi}_{PL}^{-1}\hat{\mu}_{ML,i})} = n-1.$$

The quasi-likelihood model (J. A. Nelder 2000) uses a deviance statistic rather than the Pearson statistic in the pseudo-likelihood model to estimate dispersion using a function

$$2\sum\left\{y_i \log\left[\frac{y_i}{\hat{\mu}_{ML,i}}\right] - (y_i+\hat{\phi}_{QL}^{-1})\log\left[\frac{y_i+\hat{\phi}_{QL}^{-1}}{\hat{\mu}_{ML,i}+\hat{\phi}_{QL}^{-1}}\right]\right\} = n-1.$$

Nelder and Lee (1992) found that the variance parameter from the quasi-likelihood model is more efficient than the parameter from pseudo-likelihood model(J A Nelder and Lee 1992).

1.3 Logistic regression

When the response variable is binary, binomial regression is commonly used to model the probability of an event using the inverse of a link function (g^{-1}) to a linear combination of predictors. The logit link function is widely adopted in social, genetic, epidemiologic studies following the model,

$$\Pr(y_i) = \Pr(y_i = 1|\boldsymbol{x}_i) = \frac{1}{1+\exp(-\boldsymbol{x}_i\beta^*)},$$

where β^* is a coefficient vector and $\boldsymbol{x}_i$ is i^{th} row of a design matrix.

This model fits to a generalized linear model, and the likelihood function is

$$\Pr(y|\beta^*) = L(\beta^*|y) = \prod_{i=1}^{n}\left[\left(\frac{1}{1+\exp(-\boldsymbol{x}_i\beta^*)}\right)^{y_i}\left(1-\frac{1}{1+\exp(-\boldsymbol{x}_i\beta^*)}\right)^{1-y_i}\right].$$

When the likelihood does not have a maximum, the numerical procedure provides an unstable erroneous finite value. This non-existing maximum of likelihood is often found in the case of separation. Complete separation occurs when a linear combination of predictors perfectly predicts the response variable, and quasi-complete separation occurs when data is close to complete separation or one factor in the response variable is completely predicted (Albert and Aanderson 1984). Complete or quasi-complete separation is easily found in studies having a small sample size. Although this separating predictor must be strongly associated with response variable, due to infinite coefficient and standard error estimates, the inferences could lead to inappropriate conclusions. (Zorn 2005).

An alternative approach that provides stable estimates was proposed by Firth (Firth 1993). This method removes first order bias from maximum likelihood estimates through including a small bias term in the likelihood function.

$$\log L^* \left(\beta^* | y\right) = \log L\left(\beta^* | y\right) + \frac{1}{2}\log \ \left|I\left(\beta^*\right)\right|$$

where $I(\beta^*)$ is the Fisher information matrix. This penalized likelihood approach is equivalent to a Bayesian approach with a Jeffrey's invariant prior in exponential family models. Although this method was developed to reduce small sample bias, the method performs well when the data display separation (Heinze 2006).

Gelman et al. (2008) also proposed an alternative method in a Bayesian framework. They suggested standardizing non-binary variables having a mean of 0 and a standard deviation of 0.5 and a centering binary variable with a mean of 0 and range of 1. Then independent Student-t priors, called weakly informative priors, are placed on the coefficients. The student-t priors are recommended because flat-tailed distributions enable for robust inference(Berger and Berliner 1986). Specifically, Cauchy (0, 2.5) priors are suggested as a default choice followed by the principle of weakly informative prior distributions. These priors appropriately estimate coefficients, even when separation appears in the data (Gelman et al. 2008).

1.4 Covariate Analysis

When analyzing genetic or genomic association studies, deciding whether to include covariates and which covariates to include in a model is an important consideration. Genetic studies often are structured to predict a trait from genetic variants, meaning that genetic variants are predictors and hence, variables of interest. A sample model is shown in Model 1.A

$$\text{Model 1. A: } g\big(E(Y_i)\big) = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{i2}$$

where g is a link function, $X_{ij}$is jth genetic variant of sample i $(j = 1 \ldots m)$, and X_{i2} is a covariate of sample i. The analysis is conducted for each genetic variant (m times). The same covariate, X_{i2}, is analyzed with each genetic variant because of a relationship between the covariate and the response variable, Y_i. However, the relationships between each genetic variant and the covariate are not known. In genomic studies, such as a case-control study, genomic expression values are modeled as a function of a case-control status. The model is

$$\text{Model 1. B: } g\left(E\left(X_{ij}^*\right)\right) = \beta_0^* + \beta_1^* Y_i + \beta_2^* X_{i2}.$$

where X_{ij}^* is the jth gene of sample i $(j = 1 \ldots m^*)$. Case-control status, Y_i, is the variable of interest. The analysis is repeated m^* times with different response variables. The relationship between the covariate, X_{i2}, and the case-control status, Y_i, is known, but the relationships between each gene, X_{ij}^*, and the covariate are unknown.

When the response variable is continuous, a covariate that is not associated with the variable of interest but associated with response variable, called a non-confounding predictive (NCP) covariate, often increases precision of the variable of interest, because NCP covariates explains some variability of a trait (L. D. Robinson and Jewell 1991). Such NCP covariates are commonly found in studies using Model 1.A. However, when the response variable is binary, including NCP covariates in a model can reduce power to detect associations(L. D. Robinson and Jewell 1991; Pirinen, Donnelly, and Spencer 2012). Pirinen et al. argued that the reduced power is caused by ascertainment of samples(Pirinen, Donnelly, and Spencer 2012). In the presence of correlation in samples, they showed that omitting covariates could improve the power.

A new approach was suggested by Zaitlen et al. (2012) to improve the power in ascertained case-control design. This new method estimates the parameters of a liability model utilizing externally identified information between a binary trait and covariates. Then, this method tests association between a genetic variant and residuals of the liability model (Zaitlen et al. 2012). Because these estimated effects of covariates are independent from the case-control data, this approach prevents the loss of power from ascertained covariates.

1.5 Multiple testing corrections

Multiple testing corrections are a crucial procedure when multiple hypotheses are tested simultaneously. These methods are important in genetic or genomic studies, where the number of tests may range from tens of thousands to several millions. As the number of tests dramatically increases, the importance of controlling Type-I errors also increases. One approach to handle Type-I error is to control the family-wise error rate (FWER), defined as

$$\mathrm{FWER} = \mathrm{P}(V \geq 1),$$

where V is the number of Type-I errors. In other words, it is the probability of one or more Type-I errors among a family of hypothesis tests. Another approach to handle Type-I error is controlling the false positive rate (FDR), defined as

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R} | R > 0\right) \mathrm{P}(R > 0),$$

where R is the number of rejected hypotheses (Benjamini and Hochberg 1995). FDR is developed to control the expected proportion of Type-I errors among rejected hypotheses. Because FDR is less stringent in controlling Type-I errors compared to FWER, FDR is more powerful than FWER but allows increased Type-I errors.

Among multiple testing correction methods assumptions about the dependence structure of p-values under the null hypotheses may vary. Statistical power is generally greater for those methods with stronger assumptions. P-values from alternative hypotheses are not involved in this dependence assumption.

Multiple testing methods that do not make any assumptions about the dependency structure of p-values utilize the Bonferroni’s or Hommel’s inequalities(Galambos 1977; Hommel 1986). These methods are applicable to p-values even when there is correlation among the tests performed (and hence the p-values) under the null hypothesis. Some multiple correction methods assume Positive Dependence through Stochastic Ordering, also known as the Positive Regression Dependence on Subset. This assumption allows independent or positively dependent p-values of null hypotheses. Some methods are only valid under the assumption of independent p-values, and this independence assumption is the strongest assumption.

Among multiple testing correction methods, one needs to consider whether a method assumes dependency of p-values before determining if a method is appropriate for a particular data set. Because dependency structures often exist in high-dimensional data such as genetic and genomic data, appropriate selection of a multiple testing correction method is necessary.

1.6 Dissertation outline

In this dissertation, we investigate alternative analysis methods and evaluate important aspects in RNA-Seq studies. Our research focuses on statistical inference methods including negative binomial and logistic regressions, covariate adjustment, and multiple testing methods. Each topic describes limitations of current methods, effects of those limitations, and an alternative method of overcoming those limitations that is evaluated through comprehensive simulations and a real data application.

In Chapter 2, we suggest an alternative regression method for differential expression studies using RNA-Seq data. This method simplifies the analysis procedures and removes non-biological assumptions required by conventional methods. We expect this alternative approach to reduce complexities presented in RNA-Seq studies while maintaining an appropriate Type-I error rate and power comparable to current methods.

In Chapter 3, we investigate the effect of non-predictive covariates in negative binomial regression. We expect this investigation of non-predictive covariates to demonstrate that researchers should be cautious about selecting covariates to include in statistical models for RNA-Seq data. However, this effect of non-predictive covariates in negative binomial regression is not limited to RNA-Seq studies.

In Chapter 4, we explore multiple testing correction methods specific to the analysis of RNA-Seq data. The independence assumption in some multiple testing correction methods precludes application to correlated data. The goal of this investigation is to identify a suitable multiple testing method for correlated count data, such as RNA-Seq data.

In Chapter 5, we summarize our conclusions and recommendations, and provide future directions.

Chapter 2 Evaluation of Logistic Regression Models for Case-Control Study in RNA-Seq Analysis

2.1 Introduction

Recent Next Generation Sequencing (NGS) technologies generate discrete counts of RNA sequencing (RNA-Seq). Several characteristics of RNA-Seq count data are important to account for in statistical analysis. The count of a particular gene could range from zero to several thousand, and is frequently not normally distributed. The initial RNA-Seq studies assumed the count data follow Poisson distributions(Marioni et al. 2008; Mortazavi et al. 2008; Jiang and Wong 2009). However, Poisson models cannot appropriately explain biologic dispersions of genes because the mean is equal to the variance in Poisson models. The Negative Binomial (NB) distribution more appropriately models the biological dispersion of a gene, and this NB model has been generally taken to analyze RNA-Seq data. Additionally, the total number of read counts can differ for each sample, making an appropriate normalization of RNA-Seq data necessary prior to statistical analysis of associations between status of samples (e.g. disease or not diseased) and expression level of genes.

Even if the normalization issue is addressed by applying an appropriate normalization method, the estimation of the dispersion parameter (ϕ) of each gene is very challenging with the small number of observations typically available in RNA-Seq studies. An overestimated dispersion may result in loss of power to detect differently expressed genes and an underestimated dispersion parameter may increase false discoveries. Many methods have been developed to effectively estimate the dispersion parameters, including Quasi-Likelihood (QL)(Si and Liu 2013), Weighted Quantile-Adjusted Conditional Maximum Likelihood(M. D. Robinson and Smyth 2007; M. D. Robinson, McCarthy, and Smyth 2010), Cox-Reid Adjusted Profile Likelihood(McCarthy, Chen, and Smyth 2012), and Empirical Bayes Shrinkage(Landau and Liu 2013; Love, Huber, and Anders 2014; Wu, Wang, and Wu 2013) methods. Landau and Liu reported that the selection of the estimation method may impact the test performance(Landau and Liu 2013). Two of the most sophisticated and widely used software packages for identifying differently expressed genes are DESeq2 and edgeR(Love, Huber, and Anders 2014; M. D. Robinson, McCarthy, and Smyth 2010). These two software packages estimate dispersion parameter of each gene using Empirical Bayes Shrinkage and Cox-Reid Adjusted Profile Likelihood methods, respectively.

Although NB regression has been generally accepted in the analysis of RNA-Seq data, its appropriateness in this setting has not been exhaustively evaluated. Furthermore, computational and mathematical complexity and an absence of consensus concerning appropriate methods challenges researchers conducting RNA-Seq studies(Landau and Liu 2013; Soneson and Delorenzi 2013). Because many RNA-Seq studies are designed to compare cases and controls, we explore logistic regression as an alternative approach, in which disease status is modeled as a function of RNA-Seq reads. Logistic regression is a standard method in the context of Genome-Wide Association Studies (GWAS) of binary traits. Execution of logistic regression becomes possible through reversing the experimental and explanatory variables in the NB model in the RNA-Seq setting. An attractive feature of the logistic framework in the application to RNA-Seq data is that the estimation of a dispersion parameter for gene expression is not necessary.

In this chapter, we investigate this alternative approach. We reverse the dependent variable and independent variable specified in a NB model and evaluate logistic regression models in which the dependent variable is disease status and gene expression is the independent variable. Specifically, we compare NB regression, as implemented in the DESeq2 package with Classical Logistic (CL), Bayes Logistic (BL), and Firth Logistic (FL) regression approaches. We use both simulated data sets and an application to a real Huntington's disease (HD) mRNA-Seq data set.

2.2 Dispersion estimation methods in negative binomial framework

This study treats each gene as a unit; hence various gene-based scenarios are considered. Although several methods implemented in the RNA-Seq setting utilize data from across all genes to improve estimation, we did not use those methods in our gene-focused simulations. Maximum likelihood (ML) and Quasi likelihood (QL) methods (methods described in Chapter 1.2) and the true parameter value used in simulation are used in NB regressions for analysis of all simulated data. However, in our real data application, we analyzed the HD RNA-Seq data set with the DESeq2 package and analyzed the whole gene set at once. This statistical package implements the Empirical Bayes Shrinkage Estimation method to estimate gene specific dispersion and this estimate was used for all data analyses including permutation analyses.

2.3 Regression methods for analyzing RNA-Seq data

The following section describes regression methods that are used in this comparative study. RNA-Seq reads are modeled as a function of case-control status in NB models, and case-control status is modeled as a function of RNA-Seq reads in logistic models.

2.3.1 Negative binomial regression

NB regression uses the same ML fitting process that estimates the ML dispersion. This GLM framework is used by the leading software packages DESeq2 and edgeR. In the current study, GLM was implemented using the *glm(,family=negative.binomial(1/ϕ))* function in R-package “MASS” and utilized either the estimated dispersion from ML, QL, or the true dispersion value from the simulation scenario. In our real data application, the original data and permuted data sets was analyzed with DESeq2. DESeq2 incorporates the Empirical Bayes Shrinkage method to estimate effect sizes of gene expression. Because this method shrinks some large effect sizes that are not explained well by the data toward zero, the shrunken effect sizes are more reliable than the effect sizes from ML(Love, Huber, and Anders 2014).

2.3.2 Classical logistic regression

We conducted GLM in a logistic regression framework using the logit link function. The *glm(,family=binomial)* function in R was used. Because RNA-Seq studies are commonly designed for small samples, CL regression may confront the small sample bias. Also, complete separation, which often occurs when the effect size is large, may prevent utilizing CL regression when testing for differential expression in the RNA-Seq setting. If the expression values of a gene are completely or nearly completely separated between case and control groups, the ML estimation from CL regression may fail to converge. Because observing complete separation for genes may be a promising indicator of differential expression, we implemented Bayes and Firth’s logistic regressions, which overcome complete separation in the logistic framework.

2.3.3 Bayes logistic regression

Gelman et al. proposed a prior to estimate stable coefficients in a Bayesian framework, when data show separation. The proposed prior is the Cauchy distribution with center 0 and scale 2.5(Gelman et al. 2008). They demonstrated that this flat-tailed distribution has robust inference in logistic regression and is computationally efficient. The procedure is implemented by incorporating an EM algorithm into iteratively reweighted least squares. The *bayesglm* function in the R-package “arm” was used.

2.3.4 Firth’s logistic regression

The ML estimators may be biased due to the small sample size and the small total Fisher information. Firth proposed a method that eliminates first-order bias, $O(n^{-1})$, in ML estimation by introducing a bias term in the likelihood function(Firth 1993). This correction is also equivalent to penalizing likelihood function with Jeffery’s invariant prior in Bayesian framework if the target parameters follow canonical parameters of an exponential family. Heinze and Schemper demonstrated that Firth’s method is an ideal solution when the data show separation (Heinze and Schemper 2002). Firth’s method was motivated to correct the bias in case-control samples due to small sample size(Allison 2012). The *logistf* function in the R-package “logistf” was used.

2.4 Data Adaptive (DA) distribution of test statistics

The following steps describe our DA method, which re-estimates a distribution of test statistics under the null hypotheses of no association suggested by Han and Pan(Han and Pan 2010). The DA approach enables one to obtain a recalibrated distribution of test statistics because when sample size is small, the theoretical asymptotic distribution may not be appropriate. This method also avoids heavy computing burden compared to implementing permutation tests with all possible permutations.

To implement the DA approach we need to obtain a set of Wald Chi-square test statistics $\left(U^{(1)},..,U^{(m)}\right)$ from m number of null data sets. We calculate the sample mean and variance of this null test statistic as U_0 and V_0. Because $U^{(1)},..,U^{(m)}$ follow a null empirical distribution $a\chi_1 + b$,

$$E[U] = E[a\chi_1 + b] = a + b = U_0 \,,$$

$$\mathrm{var}[U] = \mathrm{var}[a\chi_1 + b] = 2(a)^2 = V_0,$$

We can solve a and b in terms of U_0 and V_0, so that

$$a = \sqrt{\frac{\mathrm{var}[U]}{2}} = \sqrt{\frac{V_0}{2}},$$

$$b = E[U] - \sqrt{\frac{\mathrm{var}[U]}{2}} = U_0 - \sqrt{\frac{V_0}{2}}.$$

Our test statistic is then compared to the null empirical distribution $a\chi_1 + b$.

2.5 Simulation study

The simulations varied various aspects of RNA-Seq data properties and study design including sample size, mean expression value (μ), log2 fold-change (l2fc), and dispersion. The performance of statistical models was evaluated through different Type-I error and power scenarios using combinations of the parameter values in Table 2.1.

Table 2.1 Parameters and their values in simulation scenarios

Parameter	Values
Design	Balanced, Unbalanced2, Unbalanced4
Number of cases	10, 25, 75, 500
Mean expression value in controls($\mu_{D=0}$)	50, 100, 1000, 10000
Dispersion	0.01, 0.01, 0.5, 1
$\log_2$ fold-change (l2fc)	0, 0.3, 0.6, 1.2, 2

Design: Balanced has the same number of cases and controls. Unbalanced2 (4) has the 2 (or 4) times more controls than cases. log2 fold-change: The l2fc equals to $\log_2\left(\frac{\text{mean expression value in cases } (\mu_{D=1})}{\text{mean expression value in controls } (\mu_{D=0})}\right)$.

2.5.1 Generation of simulated RNA-Seq data

For each scenario, the read counts (y_g) were sampled from the NB distribution with mean and dispersion as specified in in Table 2.1. We simulated 10,000 replicates per scenario using the following steps.

First, we sampled cases and controls based on the study design. Then, a gene expression value for each sample (Y_{ig}) was sampled from the NB distribution conditioning on the disease status of the sample. The l2fc determined the mean expression values in cases ($\mu_{gD=1}$) in power scenarios. When simulating under the null hypothesis (Type-I error scenarios) l2fc was equal to 0 and the mean expression value (μ_{gD}) was equal for cases and controls. We considered only the situation in which the gene is up-regulated, and assumed that the dispersion parameter was the same for cases and controls. We can write the simulation model for the RNA-Seq count as

$$Y_{ig} \sim NB(\mu_{gD}, \phi_g), \text{where } \mu_{gD} \geq 0, \phi_g \geq 0$$

where D is a binary case-control status of sample i, μ is mean expression value of gene g, $\mu_{D=1}$is the mean expression value for cases and is calculated as $2^{\text{l2fc}} \times \mu_{D=0}$

2.5.2 Analysis of simulated RNA-Seq data

The NB regression modeled gene expression values as a function of case-control status, but the logistic regressions modeled cases-control status as a function of gene expression values. We performed the NB regression with Model 2.A and performed the CL, BL, and FL regressions with Model 2.B.

$$\text{Model 2. A: } \log(\text{E}[Y]) = \beta_0 + \beta_1 D,$$

$$\text{Model 2. B: } \text{logit}(\text{E}[D]) = \beta_0^* + \beta_1^* Y.$$

The NB regression required estimation of a dispersion parameter. Three different dispersions were used in analyses: One was estimated from ML, another was estimated from QL, and the other was assigned to the true value from the simulation scenario.

Scenarios for which l2fc is zero are Type-I error studies. Otherwise, the scenarios are power studies. Type-I error rates, at significance (alpha) levels 0.05 and 0.01, were calculated based on replicates with converged results. For power studies, different Type-I error rates observed among the distinct regression methods were corrected by computing the empirical power with an empirical threshold calculated from different Type-I error scenarios.

$$\text{Type I error rate} = \frac{\text{The number of p−values<alpha levels}}{m_s^*}, \qquad (2.1)$$

$$\text{Empirical Power} = \frac{\text{The number of p−values<Empirical thresholds}}{m_s}, \qquad (2.2)$$

$$\text{Empirical threshold} = \ \mathrm{Q}^{\mathrm{th}} \text{ smallest p} - \text{value in null hypotheses,} \qquad (2.3)$$

where m_s is the number of simulations, m_s^* is the number of converged simulations, and Q is $\mathrm{alpha} \times m_s^*$

2.5.3 Cross-Validation of data adaptive method in simulated RNA-Seq data

The results from each Type-I error scenario were randomly and evenly partitioned into 10 groups. Of the 10 groups, 9 were assigned as the training set (9000) and the remaining one was assigned as the testing (1000) set. Then, the scale (a) and location (b) parameters were estimated from test statistics using the training set.

$$\chi_g \sim a_g \chi_1 + \ b_g \text{ where } \chi_g \text{ is a test statistic of scenario } g$$

The p-values were re-generated using a scale and location adjusted chi-square distribution. For all 10 combinations of testing and training set partitions, we estimated the scale and location parameter and re-computed p-values. Type-I error rates were re-calculated for all Type-I error scenarios.

2.6 Simulation result

2.6.1 Simulation Type-I error result

Type-I error rates from the simulated results of the scenarios at two alpha levels are presented in Table 2.2 and Table 2.3. The NB regressions using ML, QL and true dispersions show almost identical levels of performance as shown in Table 2.2. When the sample size is small or the dispersion is high, the NB regression shows inflated Type-I error rates but the CL and BL regressions are conservative (see Table 2.3). Large sample size and low dispersion generally yielded Type-I error rates that were close to the specified alpha levels. The increment of $\mu_{D=0}$ is not influential, as shown in Table 2.3. The FL regression performs well or presents moderate conservativeness at both alpha levels. The Type-I error rates of the FL regression are less affected by the small sample size and the large dispersion than other logistic regressions. The Type-I error rates of additional scenarios exhibit patterns that are consistent with results in Tables 2.2 and 2.3.

Table 2.2 Type-I error rates of the NB regressions with the true dispersion and ML and QL dispersions from the balanced design

			$\alpha = 0.05$			$\alpha = 0.01$		
Ncase	**mu**	**Disp**	**NB_MLD**	**NB_TD**	**NB_QLD**	**NB_MLD**	**NB_TD**	**NB_QLD**
10	50	0.01	0.066	0.067	0.066	0.021	0.020	0.020
10	50	0.1	0.070	0.071	0.071	0.019	0.020	0.019
10	50	0.5	0.080	0.080	0.080	0.027	0.027	0.027
10	50	1	0.085	0.085	0.085	0.030	0.030	0.030
10	1000	0.01	0.066	0.066	0.066	0.018	0.018	0.018
10	1000	0.1	0.068	0.068	0.068	0.021	0.021	0.021
10	1000	0.5	0.077	0.077	0.077	0.024	0.024	0.024
10	1000	1	0.094	0.094	0.094	0.032	0.032	0.032
10	10000	0.01	0.067	0.067	0.067	0.019	0.019	0.019
10	10000	0.1	0.069	0.069	0.069	0.022	0.022	0.022
10	10000	0.5	0.076	0.076	0.076	0.025	0.025	0.025
10	10000	1	0.087	0.087	0.087	0.028	0.028	0.028
25	50	0.01	0.056	0.056	0.056	0.014	0.014	0.014
25	50	0.1	0.060	0.060	0.060	0.013	0.013	0.013
25	50	0.5	0.060	0.060	0.060	0.016	0.016	0.016
25	50	1	0.061	0.061	0.061	0.017	0.017	0.017
25	1000	0.01	0.057	0.057	0.057	0.014	0.014	0.014
25	1000	0.1	0.060	0.060	0.060	0.013	0.013	0.013
25	1000	0.5	0.062	0.062	0.062	0.018	0.018	0.018
25	1000	1	0.064	0.064	0.064	0.019	0.019	0.019
25	10000	0.01	0.059	0.059	0.059	0.015	0.015	0.015
25	10000	0.1	0.055	0.055	0.055	0.011	0.011	0.011
25	10000	0.5	0.064	0.064	0.064	0.016	0.016	0.016
25	10000	1	0.065	0.065	0.065	0.016	0.016	0.016
75	50	0.01	0.051	0.051	0.051	0.012	0.012	0.012
75	50	0.1	0.053	0.053	0.053	0.012	0.012	0.012
75	50	0.5	0.050	0.050	0.050	0.011	0.011	0.011
75	50	1	0.054	0.054	0.054	0.014	0.014	0.014
75	1000	0.01	0.054	0.054	0.054	0.012	0.012	0.012
75	1000	0.1	0.051	0.051	0.051	0.011	0.011	0.011
75	1000	0.5	0.055	0.055	0.055	0.011	0.011	0.011
75	1000	1	0.056	0.056	0.056	0.013	0.013	0.013
75	10000	0.01	0.052	0.052	0.052	0.011	0.011	0.011
75	10000	0.1	0.054	0.054	0.054	0.011	0.011	0.011
75	10000	0.5	0.056	0.056	0.056	0.011	0.011	0.011
75	10000	1	0.058	0.058	0.058	0.014	0.014	0.014

Ncase: The number of cases (equal number of controls), mu: mean expression value in cases and controls, Disp: Dispersion, NB: Negative Binomial, TD: True dispersion specified in the simulation, MLD: Maximum likelihood estimated Dispersion, QLD: Quasi-likelihood estimated Dispersion

Table 2.3 Type-I error rates of the NB and logistic regressions from the balanced design

			$\alpha = 0.05$				$\alpha = 0.01$			
Ncase	**mu**	**Disp**	**NB_TD**	**CL**	**BL**	**FL**	**NB_TD**	**CL**	**BL**	**FL**
10	50	0.01	0.066	0.026	0.026	0.045	0.021	0.000	0.001	0.008
10	50	0.1	0.070	0.024	0.023	0.044	0.019	0.000	0.001	0.007
10	50	0.5	0.080	0.023	0.022	0.043	0.027	0.000	0.001	0.008
10	50	1	0.085	0.016	0.018	0.038	0.030	0.000	0.000	0.008
10	1000	0.01	0.066	0.023	0.023	0.044	0.018	0.000	0.000	0.007
10	1000	0.1	0.068	0.024	0.025	0.046	0.021	0.000	0.001	0.009
10	1000	0.5	0.077	0.019	0.020	0.041	0.024	0.000	0.000	0.007
10	1000	1	0.094	0.016	0.017	0.039	0.032	0.000	0.001	0.007
10	10000	0.01	0.067	0.024	0.023	0.044	0.019	0.000	0.000	0.008
10	10000	0.1	0.069	0.025	0.026	0.045	0.022	0.000	0.001	0.008
10	10000	0.5	0.076	0.022	0.022	0.044	0.025	0.000	0.001	0.007
10	10000	1	0.087	0.013	0.014	0.038	0.028	0.000	0.000	0.005
25	50	0.01	0.056	0.042	0.039	0.047	0.014	0.004	0.004	0.010
25	50	0.1	0.060	0.042	0.038	0.049	0.013	0.004	0.003	0.008
25	50	0.5	0.060	0.038	0.035	0.047	0.016	0.004	0.003	0.009
25	50	1	0.061	0.030	0.028	0.041	0.017	0.002	0.002	0.006
25	1000	0.01	0.057	0.044	0.040	0.049	0.014	0.005	0.004	0.011
25	1000	0.1	0.060	0.043	0.038	0.048	0.013	0.004	0.004	0.009
25	1000	0.5	0.062	0.040	0.037	0.047	0.018	0.004	0.004	0.011
25	1000	1	0.064	0.034	0.032	0.044	0.019	0.002	0.002	0.009
25	10000	0.01	0.059	0.045	0.041	0.049	0.015	0.005	0.005	0.010
25	10000	0.1	0.055	0.039	0.034	0.044	0.011	0.003	0.003	0.007
25	10000	0.5	0.064	0.039	0.036	0.046	0.016	0.004	0.003	0.008
25	10000	1	0.065	0.031	0.027	0.042	0.016	0.002	0.002	0.008
75	50	0.01	0.051	0.046	0.045	0.048	0.012	0.009	0.008	0.010
75	50	0.1	0.053	0.048	0.046	0.050	0.012	0.009	0.008	0.010
75	50	0.5	0.050	0.042	0.040	0.044	0.011	0.006	0.005	0.008
75	50	1	0.054	0.042	0.040	0.047	0.014	0.007	0.007	0.011
75	1000	0.01	0.054	0.050	0.048	0.051	0.012	0.009	0.008	0.010
75	1000	0.1	0.051	0.045	0.043	0.047	0.011	0.007	0.007	0.009
75	1000	0.5	0.055	0.045	0.043	0.048	0.011	0.007	0.006	0.009
75	1000	1	0.056	0.045	0.043	0.048	0.013	0.007	0.006	0.010
75	10000	0.01	0.052	0.047	0.046	0.049	0.011	0.009	0.008	0.010
75	10000	0.1	0.054	0.049	0.047	0.050	0.011	0.007	0.007	0.008
75	10000	0.5	0.056	0.047	0.045	0.050	0.011	0.007	0.007	0.009
75	10000	1	0.058	0.045	0.043	0.049	0.014	0.007	0.007	0.010

Ncase: The number of cases (equal number of controls), Disp: Dispersion, NB_TD: The Negative binomial regression with the true dispersion specified in simulation, CL: Classical Logistic regression, BL: Bayes Logistic regression, FL: Firth’s Logistic regression

2.6.2 DA Type-I error simulation results

In most scenarios, the DA method reduces the inflation observed with NB regressions and the deflation observed with the CL, BL, and FL regressions as presented in Table 2.4. However, when the DA method is performed with CL and BL results with small sample size, conservative results, especially with the CL model, are still exhibited at alpha level 0.01. The DA method with NB and FL regressions showed well-controlled Type-I error rates at all alpha levels even with small sample size.

Table 2.4 Type-I error rates of the NB and logistic regressions with the DA method from the balanced design

			$\alpha = 0.05$				$\alpha = 0.01$			
Ncase	**mu**	**Disp**	**NB_TD**	**CL**	**BL**	**FL**	**NB_TD**	**CL**	**BL**	**FL**
10	50	0.01	0.039	0.049	0.045	0.041	0.012	0.003	0.007	0.011
10	50	0.1	0.047	0.057	0.054	0.050	0.010	0.003	0.005	0.010
10	50	0.5	0.046	0.068	0.061	0.054	0.010	0.004	0.007	0.011
10	50	1	0.048	0.051	0.049	0.047	0.012	0.004	0.007	0.009
10	1000	0.01	0.048	0.060	0.057	0.054	0.011	0.001	0.004	0.011
10	1000	0.1	0.040	0.050	0.047	0.042	0.008	0.001	0.006	0.009
10	1000	0.5	0.039	0.052	0.049	0.045	0.008	0.003	0.006	0.006
10	1000	1	0.043	0.058	0.054	0.048	0.009	0.005	0.006	0.007
10	10000	0.01	0.054	0.068	0.065	0.059	0.014	0.003	0.009	0.014
10	10000	0.1	0.042	0.055	0.052	0.048	0.011	0.003	0.006	0.011
10	10000	0.5	0.044	0.049	0.045	0.044	0.006	0.003	0.005	0.006
10	10000	1	0.048	0.059	0.055	0.051	0.011	0.001	0.004	0.007
25	50	0.01	0.051	0.056	0.055	0.053	0.013	0.010	0.011	0.013
25	50	0.1	0.053	0.062	0.061	0.057	0.009	0.007	0.007	0.009
25	50	0.5	0.045	0.048	0.048	0.045	0.011	0.006	0.006	0.008
25	50	1	0.061	0.061	0.061	0.061	0.013	0.007	0.008	0.011
25	1000	0.01	0.056	0.061	0.061	0.058	0.017	0.013	0.014	0.016
25	1000	0.1	0.047	0.054	0.053	0.049	0.008	0.005	0.006	0.007
25	1000	0.5	0.043	0.045	0.045	0.044	0.005	0.003	0.003	0.004
25	1000	1	0.049	0.056	0.055	0.052	0.008	0.005	0.006	0.007
25	10000	0.01	0.043	0.047	0.047	0.044	0.011	0.007	0.008	0.010
25	10000	0.1	0.054	0.057	0.056	0.054	0.010	0.008	0.009	0.010
25	10000	0.5	0.049	0.055	0.055	0.051	0.008	0.004	0.005	0.007
25	10000	1	0.045	0.050	0.049	0.047	0.016	0.010	0.010	0.013
75	50	0.01	0.039	0.041	0.041	0.040	0.008	0.007	0.007	0.008
75	50	0.1	0.054	0.057	0.057	0.055	0.011	0.010	0.010	0.010
75	50	0.5	0.048	0.053	0.053	0.050	0.008	0.006	0.006	0.007
75	50	1	0.047	0.051	0.051	0.050	0.012	0.011	0.011	0.011
75	1000	0.01	0.050	0.052	0.052	0.050	0.016	0.015	0.015	0.015
75	1000	0.1	0.053	0.056	0.056	0.053	0.009	0.008	0.008	0.009
75	1000	0.5	0.055	0.057	0.057	0.055	0.012	0.010	0.010	0.012
75	1000	1	0.042	0.046	0.046	0.044	0.011	0.008	0.008	0.009
75	10000	0.01	0.061	0.064	0.064	0.062	0.013	0.012	0.012	0.012
75	10000	0.1	0.047	0.048	0.048	0.047	0.010	0.009	0.009	0.009
75	10000	0.5	0.048	0.049	0.049	0.048	0.015	0.013	0.013	0.014
75	10000	1	0.053	0.057	0.057	0.055	0.009	0.008	0.008	0.008

Ncase: The number of cases (equal number of controls), Disp: Dispersion, NB_TD: The Negative binomial regression with the true dispersion specified in simulation, CL: Classical Logistic regression, BL: Bayes Logistic regression, FL: Firth's Logistic regression

2.6.3 Empirical power simulation results

We summarize the empirical power results in Tables 2.5 - 2.9. The performance of the NB regressions with ML, QL and true dispersions are almost identical, as seen in Table 2.5. Larger sample sizes increase power for all regression methods as shown in Tables 2.6 - 2.9. The influence of mean expression in controls appears with small l2fc (Table 2.6). When sample size, l2fc, and dispersion are small, increase of mean expression in controls leads to an increase of power at both alpha levels. When l2fc is large and dispersion is small, the CL regression shows very low power as seen in Table 2.9. The NB, BL, and FL regressions gain more power with large l2fc and low dispersion. These three regression methods have comparable empirical power in all scenarios. The CL regression yields the lowest power among all methods in all scenarios.

Table 2.5 Empirical power of NB regression with the true dispersion and ML and QL Dispersions from the balanced design with l2fc of 0.3

			$\alpha = 0.05$			$\alpha = 0.01$		
Ncase	**mu**	**Disp**	**NB_MLD**	**NB_TD**	**NB_QLD**	**NB_MLD**	**NB_TD**	**NB_QLD**
10	50	0.01	0.781	0.781	0.780	0.503	0.504	0.503
10	50	0.1	0.261	0.262	0.262	0.096	0.096	0.096
10	50	0.5	0.089	0.089	0.089	0.021	0.021	0.021
10	50	1	0.075	0.074	0.074	0.014	0.014	0.014
10	1000	0.01	0.989	0.989	0.989	0.939	0.940	0.940
10	1000	0.1	0.267	0.267	0.267	0.089	0.089	0.089
10	1000	0.5	0.093	0.093	0.093	0.024	0.024	0.024
10	1000	1	0.063	0.063	0.063	0.014	0.014	0.014
10	10000	0.01	0.992	0.992	0.992	0.948	0.948	0.948
10	10000	0.1	0.285	0.285	0.285	0.102	0.102	0.102
10	10000	0.5	0.093	0.093	0.093	0.025	0.025	0.025
10	10000	1	0.073	0.073	0.073	0.019	0.019	0.019
25	50	0.01	0.992	0.992	0.992	0.962	0.962	0.962
25	50	0.1	0.581	0.581	0.581	0.351	0.351	0.351
25	50	0.5	0.169	0.169	0.169	0.054	0.054	0.054
25	50	1	0.118	0.118	0.118	0.037	0.037	0.037
25	1000	0.01	1.000	1.000	1.000	1.000	1.000	1.000
25	1000	0.1	0.614	0.614	0.614	0.366	0.366	0.366
25	1000	0.5	0.162	0.162	0.162	0.043	0.043	0.043
25	1000	1	0.107	0.107	0.107	0.025	0.025	0.025
25	10000	0.01	1.000	1.000	1.000	1.000	1.000	1.000
25	10000	0.1	0.629	0.629	0.629	0.397	0.397	0.397
25	10000	0.5	0.169	0.169	0.169	0.065	0.065	0.065
25	10000	1	0.111	0.111	0.111	0.029	0.029	0.029
75	50	0.01	1.000	1.000	1.000	1.000	1.000	1.000
75	50	0.1	0.966	0.966	0.966	0.882	0.882	0.882
75	50	0.5	0.461	0.461	0.461	0.234	0.235	0.234
75	50	1	0.259	0.259	0.259	0.088	0.088	0.088
75	1000	0.01	1.000	1.000	1.000	1.000	1.000	1.000
75	1000	0.1	0.981	0.981	0.981	0.917	0.917	0.917
75	1000	0.5	0.424	0.424	0.424	0.216	0.216	0.216
75	1000	1	0.235	0.235	0.235	0.089	0.089	0.089
75	10000	0.01	1.000	1.000	1.000	1.000	1.000	1.000
75	10000	0.1	0.981	0.981	0.981	0.920	0.920	0.920
75	10000	0.5	0.417	0.417	0.417	0.208	0.208	0.208
75	10000	1	0.238	0.238	0.238	0.086	0.086	0.086

Ncase: The number of cases (equal number of controls), mu: mean expression values in cases and controls, Disp: Dispersion, NB: Negative Binomial, TD: The dispersion is used for the sampling, MLD: Maximum likelihood estimated Dispersion, QLD: Quasi-likelihood estimated Dispersion

Table 2.6 Empirical power of NB and logistic regressions from the balanced design with l2fc equal to 0.03

			$\alpha = \mathbf{0.05}$				$\alpha = \mathbf{0.01}$			
Ncase	**Cont.mu**	**Disp**	**NB_TD**	**CL**	**BL**	**FL**	**NB_TD**	**CL**	**BL**	**FL**
10	50	0.01	0.781	0.740	0.779	0.776	0.503	0.384	0.497	0.510
10	50	0.1	0.261	0.256	0.260	0.257	0.096	0.085	0.094	0.094
10	50	0.5	0.089	0.089	0.088	0.088	0.021	0.020	0.020	0.019
10	50	1	0.075	0.070	0.072	0.071	0.014	0.011	0.012	0.012
10	1000	0.01	0.989	0.775	0.988	0.987	0.939	0.510	0.930	0.929
10	1000	0.1	0.267	0.262	0.267	0.265	0.089	0.077	0.089	0.091
10	1000	0.5	0.093	0.088	0.091	0.093	0.024	0.020	0.021	0.022
10	1000	1	0.063	0.062	0.064	0.064	0.014	0.012	0.012	0.014
10	10000	0.01	0.992	0.744	0.992	0.991	0.948	0.515	0.946	0.944
10	10000	0.1	0.285	0.275	0.279	0.280	0.102	0.083	0.099	0.102
10	10000	0.5	0.093	0.091	0.091	0.093	0.025	0.022	0.026	0.025
10	10000	1	0.073	0.070	0.071	0.071	0.019	0.018	0.019	0.020
25	50	0.01	0.992	0.992	0.992	0.992	0.962	0.963	0.962	0.962
25	50	0.1	0.581	0.579	0.580	0.581	0.351	0.339	0.341	0.346
25	50	0.5	0.169	0.171	0.172	0.170	0.054	0.048	0.050	0.053
25	50	1	0.118	0.117	0.117	0.116	0.037	0.034	0.035	0.037
25	1000	0.01	1.000	0.996	1.000	1.000	1.000	0.990	1.000	1.000
25	1000	0.1	0.614	0.609	0.610	0.611	0.366	0.372	0.370	0.370
25	1000	0.5	0.162	0.159	0.159	0.160	0.043	0.045	0.044	0.045
25	1000	1	0.107	0.104	0.104	0.106	0.025	0.023	0.023	0.025
25	10000	0.01	1.000	0.997	1.000	1.000	1.000	0.986	1.000	1.000
25	10000	0.1	0.629	0.630	0.630	0.630	0.397	0.391	0.392	0.391
25	10000	0.5	0.169	0.169	0.169	0.169	0.065	0.058	0.059	0.059
25	10000	1	0.111	0.110	0.109	0.113	0.029	0.032	0.032	0.030
75	50	0.01	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	50	0.1	0.966	0.966	0.966	0.966	0.882	0.883	0.883	0.883
75	50	0.5	0.461	0.455	0.455	0.457	0.234	0.235	0.234	0.232
75	50	1	0.259	0.253	0.253	0.255	0.088	0.083	0.083	0.083
75	1000	0.01	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	1000	0.1	0.981	0.981	0.981	0.981	0.917	0.917	0.917	0.917
75	1000	0.5	0.424	0.424	0.424	0.424	0.216	0.215	0.215	0.211
75	1000	1	0.235	0.234	0.235	0.236	0.089	0.090	0.090	0.091
75	10000	0.01	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	10000	0.1	0.981	0.980	0.980	0.980	0.920	0.920	0.920	0.921
75	10000	0.5	0.417	0.414	0.415	0.415	0.208	0.211	0.211	0.214
75	10000	1	0.238	0.239	0.239	0.239	0.086	0.087	0.087	0.091

Ncase: The number of cases (equal number of controls), Disp: Dispersion, NB_TD: The Negative binomial regression with the true dispersion specified in simulation, CL: Classical Logistic regression, BL: Bayes Logistic regression, FL: Firth's Logistic regression

Table 2.7 Empirical power of NB and logistic regressions from the balanced design with l2fc equal to 0.06

			$\alpha = $ **0.05**				$\alpha = $ **0.01**			
Ncase	**Cont.mu**	**Disp**	**NB_TD**	**CL**	**BL**	**FL**	**NB_TD**	**CL**	**BL**	**FL**
10	50	0.01	1.000	0.497	1.000	0.998	0.995	0.303	0.995	0.995
10	50	0.1	0.728	0.682	0.720	0.721	0.465	0.342	0.445	0.455
10	50	0.5	0.222	0.210	0.210	0.215	0.076	0.055	0.063	0.064
10	50	1	0.140	0.125	0.131	0.134	0.039	0.024	0.030	0.032
10	1000	0.01	1.000	0.041	1.000	0.999	1.000	0.016	1.000	0.999
10	1000	0.1	0.782	0.731	0.775	0.775	0.516	0.355	0.496	0.508
10	1000	0.5	0.240	0.215	0.225	0.231	0.082	0.058	0.069	0.076
10	1000	1	0.126	0.118	0.121	0.122	0.037	0.025	0.027	0.031
10	10000	0.01	1.000	0.030	1.000	0.999	1.000	0.011	1.000	0.999
10	10000	0.1	0.789	0.734	0.775	0.780	0.526	0.358	0.500	0.520
10	10000	0.5	0.225	0.203	0.211	0.219	0.074	0.055	0.067	0.069
10	10000	1	0.131	0.114	0.120	0.123	0.040	0.031	0.036	0.039
25	50	0.01	1.000	0.950	1.000	0.999	1.000	0.887	1.000	0.999
25	50	0.1	0.987	0.986	0.986	0.986	0.946	0.942	0.942	0.945
25	50	0.5	0.519	0.509	0.510	0.512	0.267	0.243	0.248	0.267
25	50	1	0.295	0.284	0.284	0.284	0.124	0.100	0.103	0.115
25	1000	0.01	1.000	0.258	1.000	0.997	1.000	0.138	1.000	0.997
25	1000	0.1	0.995	0.994	0.994	0.995	0.971	0.970	0.970	0.971
25	1000	0.5	0.510	0.502	0.503	0.509	0.236	0.234	0.232	0.240
25	1000	1	0.286	0.274	0.274	0.280	0.111	0.092	0.094	0.104
25	10000	0.01	1.000	0.198	1.000	0.998	1.000	0.104	1.000	0.998
25	10000	0.1	0.996	0.996	0.997	0.997	0.976	0.973	0.974	0.975
25	10000	0.5	0.520	0.510	0.512	0.514	0.290	0.265	0.269	0.275
25	10000	1	0.303	0.291	0.292	0.302	0.126	0.116	0.119	0.121
75	50	0.01	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	50	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	50	0.5	0.942	0.940	0.940	0.941	0.831	0.828	0.827	0.828
75	50	1	0.715	0.709	0.709	0.712	0.443	0.423	0.424	0.426
75	1000	0.01	1.000	0.835	1.000	0.996	1.000	0.727	1.000	0.996
75	1000	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	1000	0.5	0.947	0.945	0.945	0.946	0.842	0.840	0.840	0.838
75	1000	1	0.709	0.704	0.704	0.707	0.448	0.446	0.446	0.451
75	10000	0.01	1.000	0.763	1.000	0.995	1.000	0.634	1.000	0.995
75	10000	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	10000	0.5	0.938	0.937	0.937	0.937	0.828	0.829	0.829	0.831
75	10000	1	0.701	0.699	0.699	0.700	0.437	0.436	0.436	0.450

Ncase: The number of cases (equal number of controls), Disp: Dispersion, NB_TD: The Negative binomial regression with the true dispersion specified in simulation, CL: Classical Logistic regression, BL: Bayes Logistic regression, FL: Firth’s Logistic regression

Table 2.8 Empirical power of the NB and logistic regressions from the balanced design with l2fc equal to 1.2

			$\alpha = $ **0.05**				$\alpha = $ **0.01**			
Ncase	**Cont.mu**	**Disp**	**NB_TD**	**CL**	**BL**	**FL**	**NB_TD**	**CL**	**BL**	**FL**
10	50	0.01	1.000	0.001	1.000	1.000	1.000	0.000	1.000	1.000
10	50	0.1	0.999	0.566	0.999	0.998	0.992	0.331	0.986	0.988
10	50	0.5	0.670	0.569	0.619	0.637	0.384	0.208	0.299	0.331
10	50	1	0.401	0.323	0.349	0.367	0.161	0.084	0.110	0.133
10	1000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
10	1000	0.1	1.000	0.493	1.000	0.999	0.995	0.258	0.990	0.992
10	1000	0.5	0.696	0.584	0.638	0.656	0.398	0.219	0.318	0.358
10	1000	1	0.382	0.304	0.332	0.350	0.168	0.087	0.111	0.137
10	10000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
10	10000	0.1	1.000	0.486	1.000	0.998	0.996	0.253	0.992	0.994
10	10000	0.5	0.681	0.585	0.629	0.651	0.400	0.220	0.321	0.357
10	10000	1	0.390	0.303	0.332	0.349	0.175	0.097	0.138	0.159
25	50	0.01	1.000	0.007	1.000	1.000	1.000	0.002	1.000	1.000
25	50	0.1	1.000	0.971	1.000	1.000	1.000	0.934	1.000	1.000
25	50	0.5	0.976	0.971	0.972	0.973	0.904	0.864	0.871	0.895
25	50	1	0.810	0.781	0.785	0.792	0.593	0.498	0.510	0.566
25	1000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
25	1000	0.1	1.000	0.951	1.000	1.000	1.000	0.899	1.000	1.000
25	1000	0.5	0.979	0.975	0.976	0.977	0.895	0.865	0.869	0.891
25	1000	1	0.800	0.767	0.771	0.785	0.560	0.468	0.480	0.531
25	10000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
25	10000	0.1	1.000	0.953	1.000	1.000	1.000	0.906	1.000	1.000
25	10000	0.5	0.979	0.975	0.975	0.977	0.921	0.892	0.896	0.909
25	10000	1	0.815	0.790	0.794	0.807	0.588	0.519	0.533	0.559
75	50	0.01	1.000	0.074	1.000	0.999	1.000	0.036	1.000	0.999
75	50	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	50	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	50	1	0.999	0.998	0.998	0.999	0.990	0.986	0.987	0.988
75	1000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
75	1000	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	1000	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	1000	1	0.999	0.999	0.999	0.999	0.991	0.989	0.989	0.991
75	10000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
75	10000	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	10000	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	10000	1	0.999	0.999	0.999	0.999	0.990	0.988	0.988	0.990

Ncase: The number of cases (equal number of controls), Disp: Dispersion, NB_TD: The Negative binomial regression with the true dispersion specified in simulation, CL: Classical Logistic regression, BL: Bayes Logistic regression, FL: Firth's Logistic regression

Table 2.9 Empirical power of NB and logistic regressions from the balanced design with l2fc equal to 2

			$\alpha = 0.05$				$\alpha = 0.01$			
Ncase	**Cont.mu**	**Disp**	**NB_TD**	**CL**	**BL**	**FL**	**NB_TD**	**CL**	**BL**	**FL**
10	50	0.01	1.000	0.001	1.000	1.000	1.000	0.000	1.000	1.000
10	50	0.1	0.999	0.566	0.999	0.998	0.992	0.331	0.986	0.988
10	50	0.5	0.670	0.569	0.619	0.637	0.384	0.208	0.299	0.331
10	50	1	0.401	0.323	0.349	0.367	0.161	0.084	0.110	0.133
10	1000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
10	1000	0.1	1.000	0.493	1.000	0.999	0.995	0.258	0.990	0.992
10	1000	0.5	0.696	0.584	0.638	0.656	0.398	0.219	0.318	0.358
10	1000	1	0.382	0.304	0.332	0.350	0.168	0.087	0.111	0.137
10	10000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
10	10000	0.1	1.000	0.486	1.000	0.998	0.996	0.253	0.992	0.994
10	10000	0.5	0.681	0.585	0.629	0.651	0.400	0.220	0.321	0.357
10	10000	1	0.390	0.303	0.332	0.349	0.175	0.097	0.138	0.159
25	50	0.01	1.000	0.007	1.000	1.000	1.000	0.002	1.000	1.000
25	50	0.1	1.000	0.971	1.000	1.000	1.000	0.934	1.000	1.000
25	50	0.5	0.976	0.971	0.972	0.973	0.904	0.864	0.871	0.895
25	50	1	0.810	0.781	0.785	0.792	0.593	0.498	0.510	0.566
25	1000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
25	1000	0.1	1.000	0.951	1.000	1.000	1.000	0.899	1.000	1.000
25	1000	0.5	0.979	0.975	0.976	0.977	0.895	0.865	0.869	0.891
25	1000	1	0.800	0.767	0.771	0.785	0.560	0.468	0.480	0.531
25	10000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
25	10000	0.1	1.000	0.953	1.000	1.000	1.000	0.906	1.000	1.000
25	10000	0.5	0.979	0.975	0.975	0.977	0.921	0.892	0.896	0.909
25	10000	1	0.815	0.790	0.794	0.807	0.588	0.519	0.533	0.559
75	50	0.01	1.000	0.074	1.000	0.999	1.000	0.036	1.000	0.999
75	50	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	50	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	50	1	0.999	0.998	0.998	0.999	0.990	0.986	0.987	0.988
75	1000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
75	1000	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	1000	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	1000	1	0.999	0.999	0.999	0.999	0.991	0.989	0.989	0.991
75	10000	0.01	1.000	0.000	1.000	1.000	1.000	0.000	1.000	1.000
75	10000	0.1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	10000	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
75	10000	1	0.999	0.999	0.999	0.999	0.990	0.988	0.988	0.990

Ncase: The number of cases; and the number of controls are the same, Disp: Dispersion, NB_TD: The Negative binomial regression with the dispersion used for the sampling, CL: Classical Logistic regression, BL: Bayes Logistic regression, FL: Firth’s Logistic regression

2.7 Application to RNA-Seq data of Huntington's Disease (HD)

A real RNA-Seq data set was analyzed using the DESeq2 R-package, which implements a NB generalized linear model. The data set was also analyzed outside the package utilizing R (v3.0.0) to implement CL, BL, and FL regressions. DESeq2 was not used to analyze our simulated data sets because DESeq2 was designed for analyzing a set of genes whereas our simulations focused on evaluating a scenario (gene) with a specified mean and dispersion. Although DESeq2 introduced the empirical Bayes shrinkage method for estimating dispersion and effect size, results from DESeq2 would be similar to those from the NB regression used to analyze our simulated data, if all genes in a data set come from the same distribution. The logistic regressions modeled case-control status as a function of normalized counts of a gene and covariates. The normalization was performed using DESeq2.

We examined a publicly available HD data set(Labadorf et al. 2015) downloaded from the GEO database (GSE64810). RNA was extracted from frozen brain tissue in prefrontal cortex Brodmann Area 9 from 20 HD cases and 49 controls who were neurologically normal at death and sequenced using Illumina HiSeq2000 technology for 100nucleotide paired-end reads. These reads were aligned to the human reference genome (hg19) and annotated with Gencode database (v17). Only genes that have non-zero counts in more than half of the samples were kept for analysis, and extreme outliers in the raw counts were trimmed. After filtering, there were 28,087 genes in the final data set. Age at death (AAD) categorized into 4 groups and the RNA Integrity Number (RIN) defined as a binary variable specifying RIN > 7 or <= 7) were included in the model as covariates to prevent spurious associations. Because AAD was considered a non-ordinal, categorical variable, the total number of covariates is 4 in this model. The outlier correcting method implemented in DESeq2 was not applied because the outliers were already trimmed in the raw data.

2.8 Permutation design

Permutations produce multiple null data sets from real data, and these null data sets allow us to generate a null test statistic distribution for each gene. Permutation tests compared with our alpha levels enable evaluation of Type-I error rates of each gene. This analysis allows us to assess whether the results from our simulations can be validated in real data. Permutation tests compared with results from the original HD analysis obtain exact p-values of genes. Specific details of the permutations performed are provided in Section 2.8.1.

We also applied the DA method used in our simulation studies to the real data. The test statistic distribution of each gene is re-estimated using the test statistics from permuted data sets.

2.8.1 Generation of permuted RNA-Seq data

It is important that the permuted data sets are sampled from the distribution under the null hypotheses. The following describes steps to generate a completely null permuted data set considering the effect of covariates. The original study used RIN and AAD as covariates in the model. RIN was adjusted in a model due to the potential confounding effect between HD and the abundance of RNAs. To remove the effect of RIN in our permutations, at first, samples were divided by RIN categories. Then, each gene is resampled within each category of RIN. Because AAD was included in the regression model due to its association with HD, the relationship between HD and AAD was preserved during the permutation process. We generated 10,000 Monte-Carlo permutations.

2.8.2 Analysis of permuted HD RNA-Seq data

For the original HD data and each permutated data set, DE genes between HD cases and controls were identified using the NB model (Model 2.C) as implemented in DESeq2. We also implemented the CL, BL, and FL regressions analyzing association between normalized gene counts and HD status with Model 2.D to compare statistical models.

$$\text{Model 2.C: } \log_2(E[Y]) = \beta_0 + \beta_1 D + \beta_2 AAD_{1\,vs\,2} + \beta_3 AAD_{1\,vs\,3} + \beta_4 AAD_{1\,vs\,4} + \beta_5 RIN,$$

$$\text{Model 2.D: } \text{logit}(E[D]) = \beta_0^* + \beta_1^* Y + \beta_2^* AAD_{1\,vs\,2} + \beta_3^* AAD_{1\,vs\,3} + \beta_4^* AAD_{1\,vs\,4} + \beta_5^* RIN,$$

where AAD consists of 4 groups and group 1 is the reference group.

The Type-I error rates at our alpha levels and the exact p-values(Phipson and Smyth 2010) were calculated with the results from the 10,000 permutations.

$$\text{Type I error rate} = \frac{\text{The number of p−values < alpha levels}}{m_p^*},$$

$$\text{Exact p} - \text{value } = \sum_{r_t=0}^{m_{p,t}} P(R < r|R_t = r_t)P(R_t = r_t|H_0) = \frac{\sum_{r_t=0}^{m_{p,t}} F(r;m_p,p_t)}{m_{p,t}+1},$$

where m_p is the number of permutations, m_p^* is the number of converged permutation results, R is the number of p-values less than or equal to the observed p-value(r), $R_t$is the total number of possible p-values less than or equal to the observed p-value, p_t is$(R_t + 1)/(m_{p,t} + 1)$, R assumes a binomial distribution with size of m and probability of p_t conditioning on $R_t = g_t$, and R_t follows a discrete uniform distribution on $(0, m_{p,t})$ (Phipson and Smyth 2010).

The DA method was applied using our permutation results(Han and Pan 2010) to measure Type-I error rates and to obtain adjusted p-values of each gene. The same cross-validation procedures conducted in our simulation study were applied to each gene at our alpha levels. The p-values of each gene in the original results were re-computed with scale (a) and location (b) parameters as follows

$$\chi_g \sim a_g\chi_1 + b_g, \text{where } \chi_g \text{ is the test statistic of } g^{th} \text{ gene, and } g = 1,..,28087.$$

These parameters were estimated using 1,000 randomly selected permutation results.

Asymptotic, exact, and DA p-values were corrected for multiple testing by imposing a False Discovery Rate (FDR) of 0.05. To assess the adequacy of our models, QQ-plots of original, exact and DA p-values were generated, and the genomic inflation factors, λ_{gc}, were calculated. The genomic inflation factor quantifies how closely a distribution of observed p-values is to a null distribution of expected p-values. Thus, a high genomic inflation factor may suggest evidence of inflation in the test statistics(Devlin and Roeder 1999)

2.9 Permutation result

2.9.1 Permutation Type-I error result

The Type-I error rates from the permuted data sets at two alpha levels are shown in Figure 2.1 and Table 2.10. We categorize genes into 5 groups by the estimated dispersion of a gene: (0,0.05], (0.05, 0.15], (0.15, 0.8], (0.8, 1.5], and (1.5, 10]. We may consider that a gene having an estimated dispersion parameter greater than 0.8 is largely dispersed.

In DESeq2 results, as dispersion increases, the Type-I error rates increase when genes are in the categories of the (0,0.05), (0.05, 0.15), and (0.15, 0.8). However, genes in the (0.8, 1.5), and (1.5, 10) categories exhibit decreasing Type-I error rates. Genes in the (0.8, 1.5), and (1.5, 10) categories largely have very low mean expression values. After excluding genes having mean expression values less than 3, Type-I error rates increase as the estimated dispersion increases as shown in Figure 2.1(B) and (D) and Table 2.10. These increasingly liberal Type-I error rates are observed at both alpha levels of 0.05 and 0.01, and are consistent with our simulation results.

Figure 2.1 Type-I error rates from DESeq2 analysis of the permuted HD data

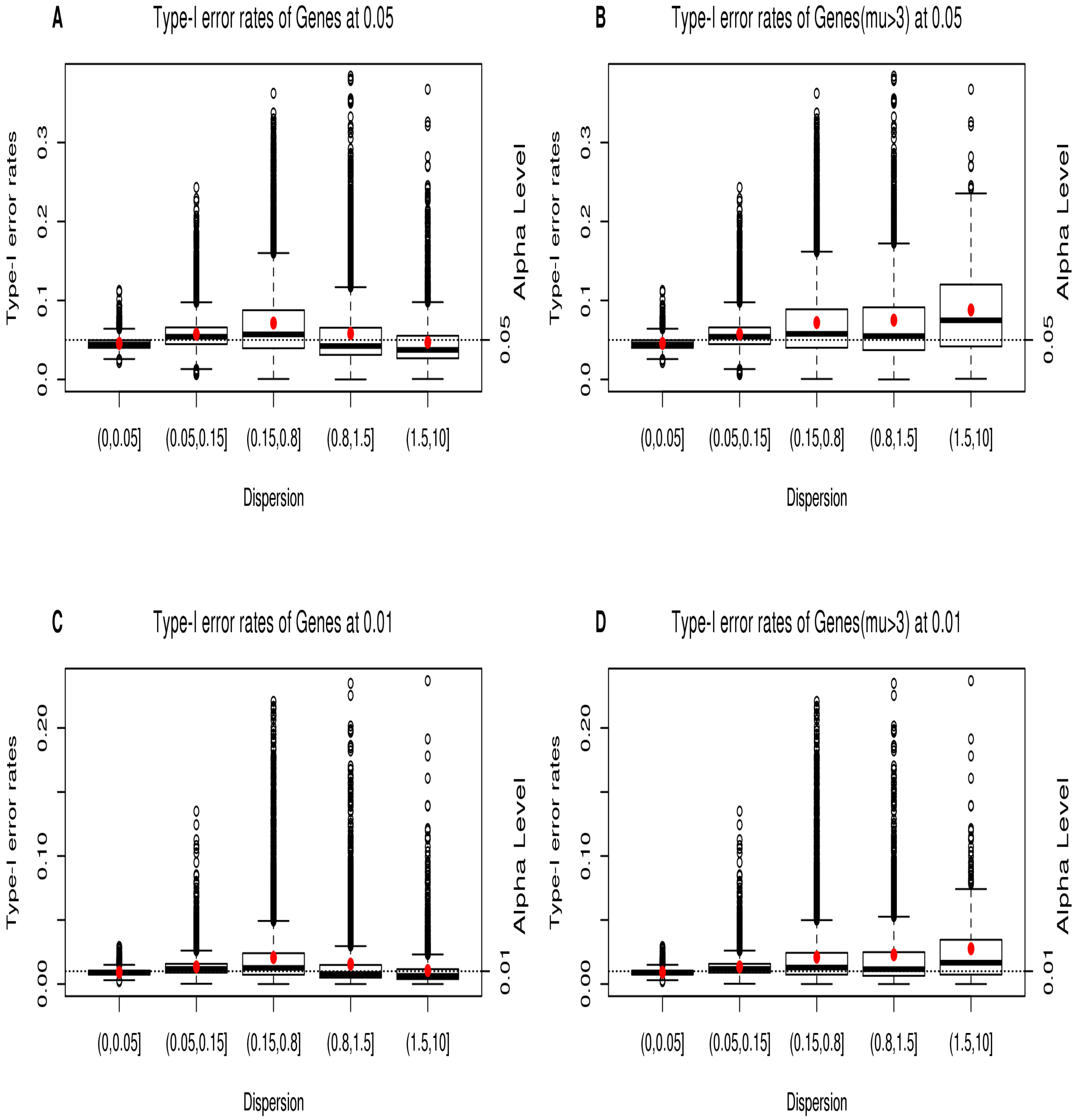


Figure 2.1 contains Type-I error rates from DESeq2 (negative binomial model) analysis of the permuted HD data at alpha levels of 0.05 and 0.01. Each black empty dot represents Type-I error rate of a gene. The red dots denote average values of Type-I error rates in each category of dispersion groups. The black dotted horizontal lines are the nominal alpha levels. Figure 2.1(A) summarizes Type-I error rates of all genes at nominal alpha level of 0.05, and Figure 2.1(B) shows Type-I error rates of genes having mean expression value of greater than 3 at alpha level of 0.05. Figure 2.1(C) represents Type-I error rates of all genes at alpha level of 0.01, and Figure 2.1(D) displays Type-I error rates of genes having mean expression value of greater than 3 at alpha level of 0.01.

Table 2.10 Type-I error rates from DESeq2 analysis of permuted HD data with mean expression value > 3

mu > 3		Dispersion Group				
		(0,0.05]	(0.05,0.15]	(0.15,0.8]	(0.8,1.5]	(1.5,10]
0.05	**Mean**	0.046	0.057	0.072	0.075	0.088
	Sd	0.009	0.021	0.048	0.058	0.062
0.01	**Mean**	0.009	0.013	0.021	0.023	0.028
	Sd	0.003	0.008	0.024	0.030	0.032

mu: mean expression values of all samples, 0.05 and 0.01: Significant levels, Sd: Standard Deviation.

In the CL, BL and FL regression results, we observe that genes in the categories of (0,0.05), (0.05, 0.15), and (0.15, 0.8) produce increasingly conservative Type-I error rates at both alpha levels, as presented in Figure 2.2 and Table 2.11. However, these increasingly conservative Type-I error rates are attenuated in the (0.8, 1.5), and (1.5, 10) categories. Because we observe this inconsistent pattern of Type-I error rates among extremely lowly expressed genes in the DESeq2 results, we also examined the set of genes excluding those with mean expression values less than or equal to 3. After exclusion, the remaining genes show consistent increasingly conservative Type-I error rates as dispersion increases as shown in Figure 2.2(B) and (D) and Table 2.11. Although Type-I error rates from the FL regression also shows more conservative when dispersion is large, Type-I error rates are relatively well controlled at both alpha levels compared to CL and BL regressions. The Type-I error rates observed in the real data set using logistic regression confirm our simulation results.

Figure 2.2 Type-I error rates from logistic models of the permuted HD data

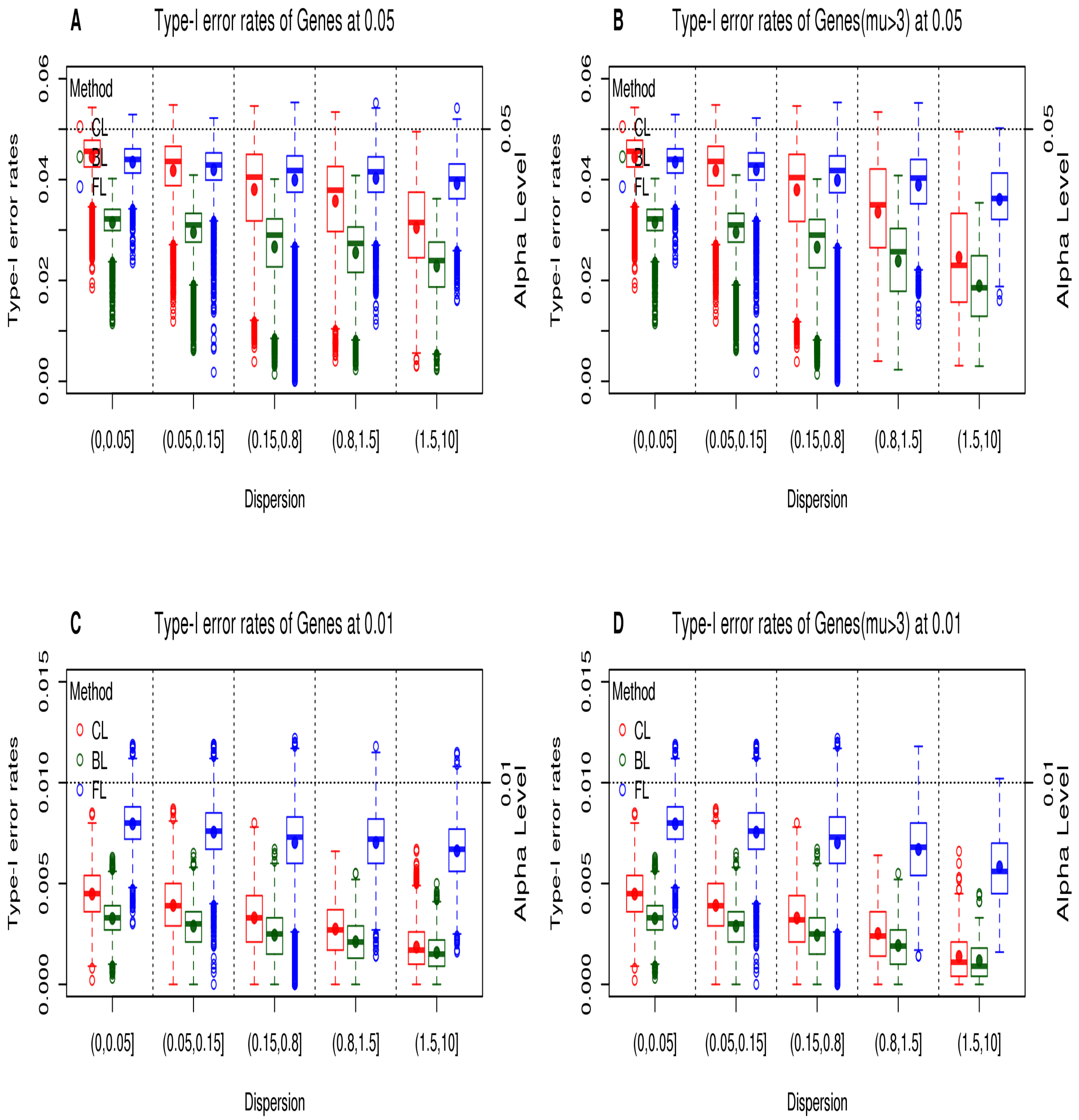


Figure 2.2 contains Type-I error rates from Classical Logistic (CL), Bayes Logistic (BL), Firth's Logistic (FL) regressions of the permuted HD data at alpha levels of 0.05 and 0.01. Each empty dot represents Type-I error rate of a gene. The dots filled with colors inside of boxes denote average values of Type-I error rates in each category of dispersion groups. The black dotted horizontal lines are our alpha levels. Figure 2.2(A) summarizes Type-I error rates of all genes at alpha level of 0.05, and Figure 2.2(B) shows Type-I error rates of genes having mean expression value of greater than 3 at alpha level of 0.05. Figure 2.2(C) represents Type-I error rates of all genes at alpha level of 0.01, and Figure 2.2(D) displays Type-I error rates of genes having mean expression value of greater than 3 at alpha level of 0.01.

Table 2.11 Type-I error rates from CL, BL, FL regressions of permuted HD data with mean expression value > 3

mu > 3		**Dispersion Group**				
		(0,0.05]	**(0.05,0.15]**	**(0.15,0.8]**	**(0.8,1.5]**	**(1.5,10]**
0.05	**Mean.CL**	0.044	0.042	0.038	0.034	0.025
	Sd.CL	0.005	0.007	0.009	0.010	0.011
	Mean.BL	0.031	0.030	0.027	0.024	0.019
	Sd.BL	0.004	0.006	0.007	0.008	0.007
	Mean.FL	0.043	0.042	0.040	0.039	0.036
	Sd.FL	0.004	0.005	0.007	0.007	0.007
0.01	**Mean.CL**	0.004	0.004	0.003	0.003	0.001
	Sd.CL	0.001	0.001	0.002	0.001	0.001
	Mean.BL	0.003	0.003	0.002	0.002	0.001
	Sd.BL	0.001	0.001	0.001	0.001	0.001
	Mean.FL	0.008	0.008	0.007	0.007	0.006
	Sd.FL	0.001	0.001	0.002	0.002	0.002

mu: mean expression values of all samples, 0.05 and 0.01: Significant levels, Sd: Standard Deviation. CL: Classical Logistic regression, BL: Bayes Logistic, FL: Firth's Logistic

2.9.2 Permutation DA method Type-I error result

The DA method controls Type-I error rates well for the DESeq2 results (Figure 2.3 and Table 2.12) and the FL regression results (Figure 2.4 and Table 2.13) at both alpha levels, regardless of dispersions of all genes. Although the Type-I error rates are well controlled in the results from CL and BL regressions at significance level of 0.05, the Type-I error rates at significance level of 0.01 are conservative as seen in Figure 2.4(B) and Table 2.13.

Figure 2.3 Type-I error rates from DESeq2 analysis with the DA method from the permuted HD data

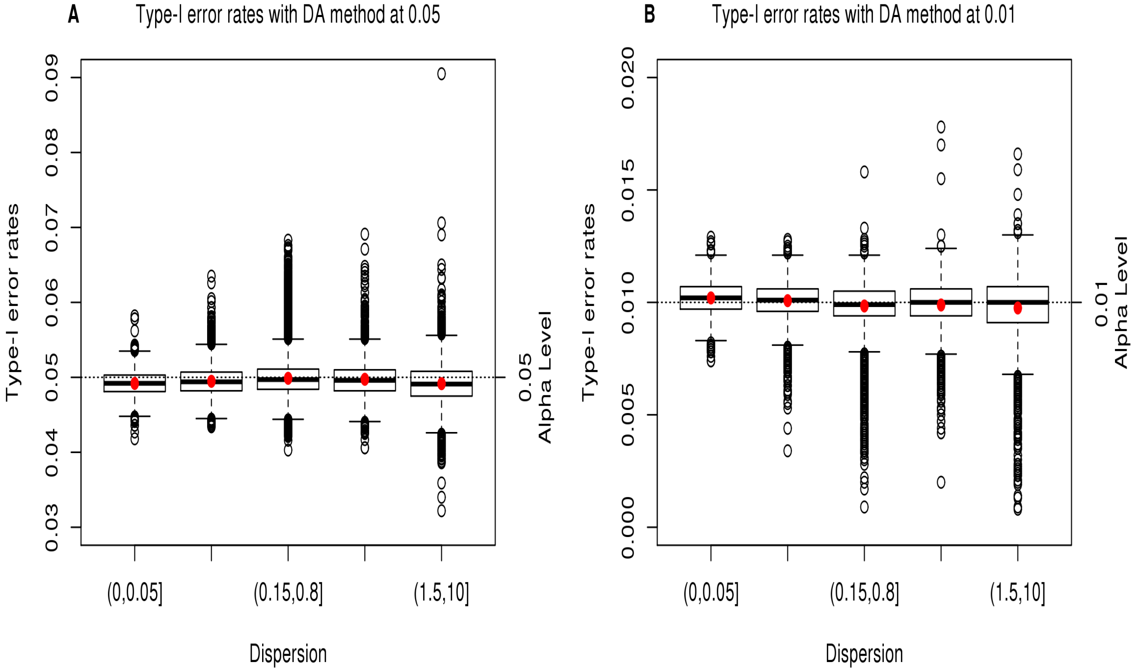


Figure 2.3 contains Type-I error rates from DESeq2 (negative binomial model) analysis with DA method of the permuted HD data at alpha levels of 0.05 and 0.01. Each black empty dot represents Type-I error rate of a gene. The red dots denote average values of Type-I error rates in each category of dispersion groups. The black dotted horizontal lines are our alpha levels. Figure 2.3(A) summarizes Type-I error rates of all genes with DA method at alpha level of 0.05. Figure 2.3(B) displays Type-I error rates of all genes with DA method at alpha level of 0.01.

Table 2.12 Type-I error rates from DESeq2 analysis with the DA method from the permuted HD data

DA		Dispersion Group				
		(0,0.05]	(0.05,0.15]	(0.15,0.8]	(0.8,1.5]	(1.5,10]
0.05	Mean	0.049	0.049	0.050	0.050	0.049
	Sd	0.002	0.002	0.002	0.002	0.003
0.01	Mean	0.010	0.010	0.010	0.010	0.010
	Sd	0.001	0.001	0.001	0.001	0.001

DA: Data Adaptive Method, 0.05 and 0.01: Significant levels, Sd: Standard Deviation.

Figure 2.4 Type-I error rates from logistic models with the DA method from the permuted HD data

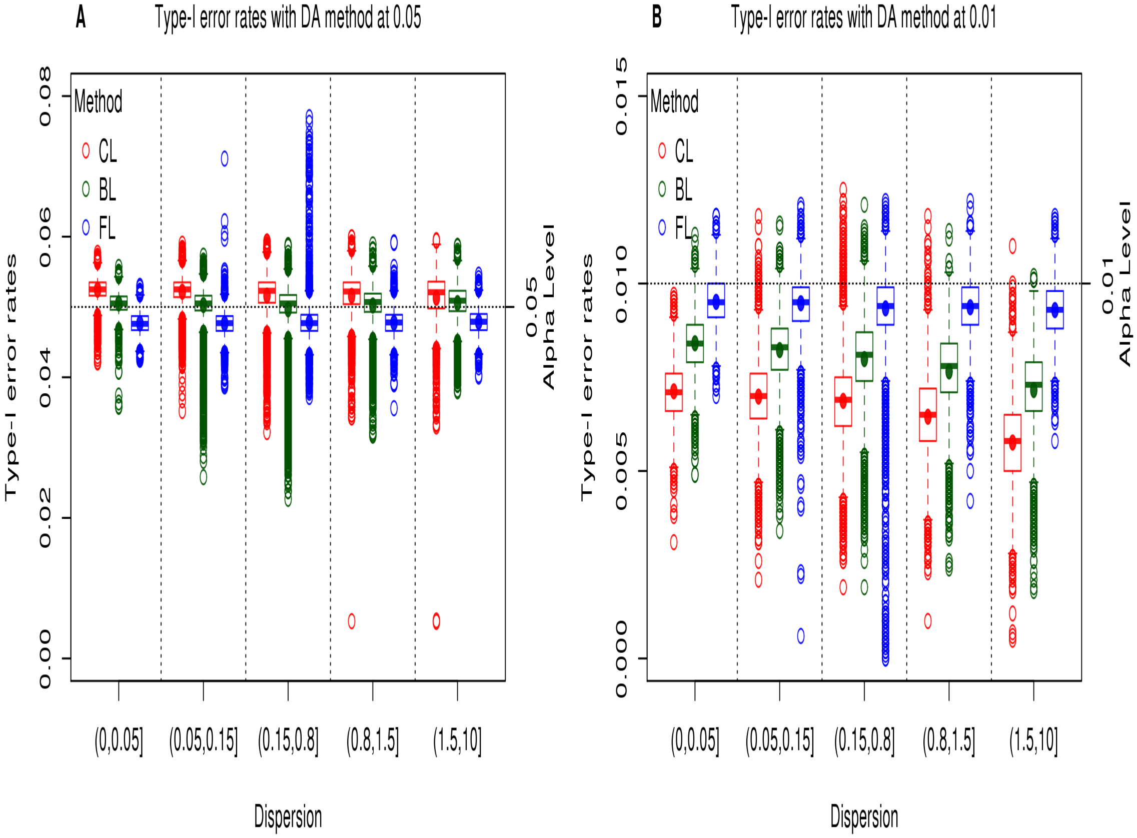


Figure 2.4 presents Type-I error rates from Classical Logistic (CL), Bayes Logistic (BL), Firth's Logistic (FL) regressions with the DA method of the permuted HD data at alpha levels of 0.05 and 0.01. Each empty dot represents Type-I error rate of a gene. The dots filled with colors inside of boxes denote average values of Type-I error rates in each category of dispersion groups. The black dotted horizontal lines are our alpha levels. Figure 2.4(A) shows Type-I error rates of all genes with DA method at alpha level of 0.05. Figure 2.4(B) represents Type-I error rates of all genes with DA method at alpha level of 0.01.

Table 2.13 Type-I error rates from logistic models with the DA method from the permuted HD data

DA		Dispersion Group				
		(0,0.05]	**(0.05,0.15]**	**(0.15,0.8]**	**(0.8,1.5]**	**(1.5,10]**
0.05	**Mean.CL**	0.052	0.052	0.052	0.052	0.051
	Sd.CL	0.002	0.002	0.003	0.003	0.004
	Mean.BL	0.051	0.050	0.050	0.050	0.051
	Sd.BL	0.002	0.003	0.004	0.003	0.003
	Mean.FL	0.048	0.048	0.048	0.048	0.048
	Sd.FL	0.002	0.002	0.002	0.002	0.002
0.01	**Mean.CL**	0.007	0.007	0.007	0.006	0.006
	Sd.CL	0.001	0.001	0.001	0.001	0.001
	Mean.BL	0.008	0.008	0.008	0.008	0.007
	Sd.BL	0.001	0.001	0.001	0.001	0.001
	Mean.FL	0.010	0.009	0.009	0.009	0.009
	Sd.FL	0.001	0.001	0.001	0.001	0.001

DA: Data Adaptive Method, mu: mean expression values of all samples 0.05 and 0.01: Significant levels, Sd: Standard Deviation, CL: Classical Logistic regression, BL: Bayes Logistic, FL: Firth's Logistic

2.9.3 HD RNA-Seq data analysis results

We analyze the HD data using NB GLM in the DESeq2 R-package, and analyze the data using CL, BL, and FL regressions also in R functions described in Sections 2.3.2 – 2.3.4. All regression results are corrected with the DA method, and are adjusted for multiple testing using an FDR of 0.05. The Q-Q plots and genomic control lambdas are shown in Figure 2.5. The DA method reduced the mean of the lambdas from the results of DESeq2 and increased the mean of the lambdas from the results of the CL, BL, and FL regressions. As shown in Figure 2.6, we identified 3,203 genes that were significant across all methods. The FL regression also identified 307 genes as differentially expressed that were not identified by the other methods. The DESeq2 approach identified 944 genes that were not identified as significant using the other methods. Of the genes that are not significant (FDR > 0.05) in the DESeq2 analysis but significant (FDR < 0.05) in CL, BL, FL regressions, the 10 most significant (FDR < 0.05) from the FL regression are shown in Table 2.14. The most significant gene is *SLC1A6* with p-values 3.17E-06 from the FL regression, respectively. Of the genes that are not significant (FDR > 0.05) in the CL, BL and FL analyses, the 10 most significant (FDR < 0.05) from DESeq2 are shown in Table 2.15.

Figure 2.5 Q-Q plots of HD data analysis by regression methods

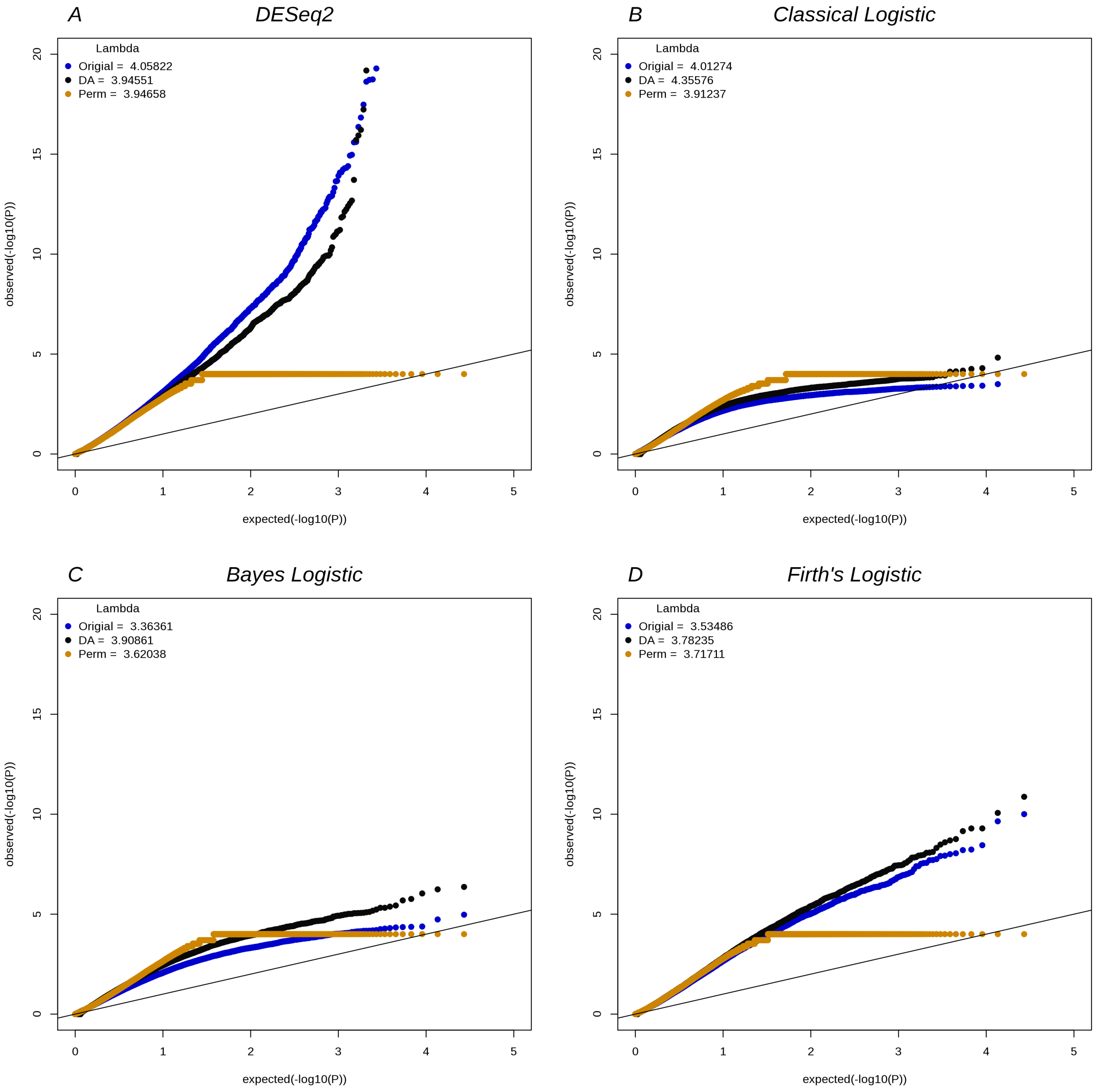
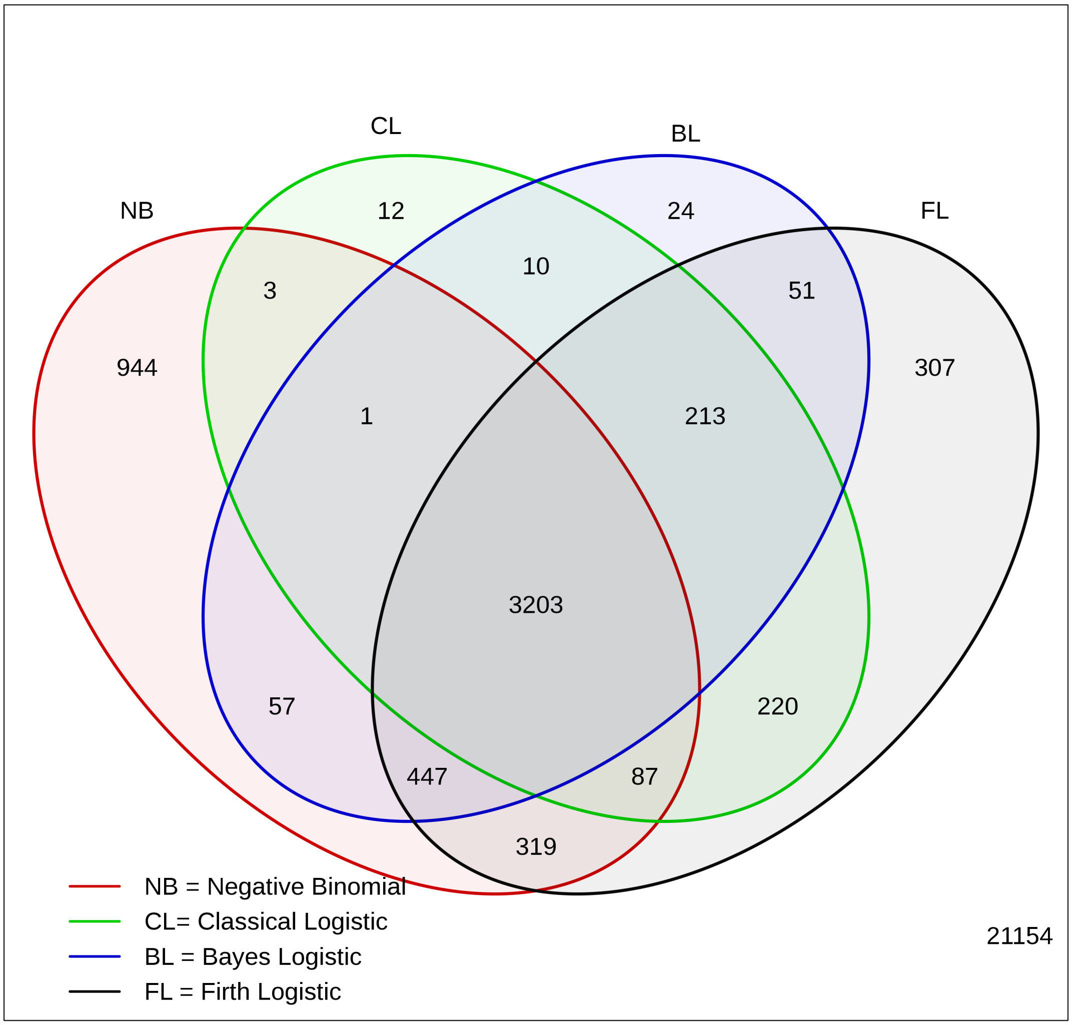


Figure 2.5 exhibits the Q-Q plots from the HD analysis adjusting for age at death and RIN from DESeq2 (A), and Classical (B), Bayes (C), and Firth’s (D) Logistic regressions. Each regression method contains three different ways of calculating p-values (Original, DA, and Perm). “Original” p-values (Blue dots) are estimated from theoretical asymptotic distribution. “DA” p-values (Black dots) are evaluated from data adaptive asymptotic distribution using 1,000 permutations. “Perm” p-values (Yellow dots) are calculated using 10,000 permutations.

Figure 2.6 Venn diagram of HD analysis results using DA method



Each colored circle represents a different regression method. The numbers inside of the circles are the number of genes significant at FDR 0.05 based on p-values adjusted using the Data Adaptive (DA) method. There were 3,203 significant genes in common across all the methods. The FL identified the largest number of significant genes compared to CL and BL. The NB independently identified 944 genes.

Table 2.14 Top 10 genes from FL regressions among genes having FDR > 0.05 in DESeq2 and FDR < 0.05 in CL, BL, and FL regressions using the DA method

Gene	Case	Cont	Disp	NB	CL	BL	FL
SLC1A6	373.5	553.9	0.29	0.039	4.33E-04	4.52E-04	3.17E-06
SERHL2	209.7	163.4	0.17	0.016	3.17E-04	6.26E-03	1.15E-05
KCNK9	314.6	453.9	0.30	0.063	3.02E-04	9.22E-04	1.72E-05
DISP2	686.7	936.6	0.21	0.047	5.54E-04	8.17E-04	4.25E-05
SPOCK2	12370.2	15648.9	0.09	0.010	8.87E-04	1.05E-03	8.04E-05
C20orf27	726.0	933.5	0.11	0.019	5.91E-04	2.44E-04	9.50E-05
IST1	3387.8	3133.6	0.02	0.009	5.68E-04	4.54E-03	9.59E-05
ARC	595.8	1058.2	0.40	0.030	1.06E-03	1.15E-03	1.03E-04
STRADB	980.3	844.0	0.03	0.013	1.36E-03	1.54E-03	1.07E-04
PCP4	734.3	1329.5	0.37	0.086	1.09E-03	2.57E-03	1.15E-04

Case: Normalized mean expression value in cases, Cont: Normalized mean expression value in controls, Disp: Dispersion, NB: P-values from negative binomial regression with true dispersion, CL: P-values from classical logistic regression, BL: P-values from Bayes logistic regression, FL: P-values from Firth's logistic regression.

Table 2.15 Top genes from DESeq2 among genes having FDR > 0.05 in CL, BL and FL regressions using the DA method

Gene	Case	Cont	Disp	NB	CL	BL	FL
RP11-115J23.1	2.8	0.4	2.19	9.67E-06	0.019	0.011	0.012
CTD-2281E23.3	0.6	2.5	1.20	3.26E-05	0.028	0.016	0.020
LL22NC03-104C7.1	1.1	7.3	1.54	3.42E-05	0.036	0.014	0.009
CEACAM3	2.9	0.4	2.53	5.62E-05	0.044	0.020	0.016
RP11-351I21.6	1.7	11.8	1.58	6.72E-05	0.043	0.022	0.023
LINC00310	29.7	9.9	0.73	9.91E-05	0.025	0.013	0.010
RP5-850O15.3	0.4	2.8	1.55	1.06E-04	0.020	0.018	0.010
RP11-554A11.9	15.4	37.8	0.55	1.70E-04	0.014	0.009	0.009
GK3P	8.1	15.9	0.65	1.75E-04	0.020	0.013	0.014
S100A11	568.0	266.0	0.44	2.43E-04	0.019	0.015	0.013

Case: Normalized mean expression value in cases, Cont: Normalized mean expression value in controls, Disp: Dispersion, NB: P-values from negative binomial regression with true dispersion, CL: P-values from classical logistic regression, BL: P-values from Bayes logistic regression, FL: P-values from Firth's logistic regression.

2.10 Discussion

We propose using a logistic regression framework as an alternative to Negative Binomial (NB) regression to analyze RNA-Seq data for case-control studies. We have shown in our simulations that Firth Logistic (FL) regression performs well in terms of controlling Type-I error rates and shows comparable empirical power. The dispersion is not estimated in the logistic framework, thus avoids potential false association resulting from incorrectly estimated dispersions, and is statistically succinct. Because the Bayes Logistic (BL) and FL regressions overcomes complete separation, the empirical power for these methods are very close to the power observed for NB regression in contrast to classic logistic (CL). The simulations presented focused on single genes varying relevant parameters (mean, dispersion, log fold change); transcriptome-wide data was not simulated.

The Type-I error simulations presented demonstrate that NB regression has inflated Type-I error rates, and Classical Logistic (CL) and BL regressions are very conservative with small sample size. The degrees of inflation/deflation varied by the scale of the dispersion parameter within the same sample size. This variation by the dispersion parameter is confirmed through the observed Type-I error rates from permutation of a real data set. Although large sample size could reduce the inflation from NB and the deflation from CL and BL regressions, the high cost of RNA-Seq technology and difficulty of obtaining certain sample tissues, such as human brain, may preclude a larger sample size in some studies. The distinct Type-I error rates observed with varying dispersion parameter values may violate the general assumption that p-values from non-DE genes follow a uniform distribution. However, the current simulation and permutation studies validate that the DA (Data Adaptive) method is a suitable alternative approach that controls Type-I error rates in all regression methods.

The empirical power of the NB, BL, and FL regressions are comparable across all scenarios. Lower power was observed for CL regression, which appears to be driven by scenarios of complete separation and a failure of CL models to converge. When simulation scenarios have large l2fc and small dispersion, simulated data are likely to show complete separation. The NB, BL and FL regressions are powerful in these scenarios. In most scenarios, the CL regression demonstrated the lowest empirical power among all methods.

Unlike NB, CL and BL regressions, FL regression controls Type-I error rates well and maintains comparable power even with small sample size. Firth logistic regression is an excellent alternative to NB regression for analysis of RNA-Seq data in case-control studies.

Analysis of the HD data showed the genomic inflation factor was decreased after applying the DA method to the results from NB GLM but the genomic lambdas were increased after applying DA method to the results from CL, BL and FL regression models. The exact p-values from 10,000 permutations revealed the same pattern. This pattern is consistent with our simulation results where we observed inflated Type-I error rates in the NB framework and deflated in the logistic framework when test statistics were compared with a theoretical asymptotic distribution.

Although it is unknown which genes are truly differentially expressed in the HD data set, we compared DE genes identified in the HD data by different statistical approaches. We found that *SLC1A6* (solute carrier family 1, member 6; *EAAT4*) did not show evidence of association with HD when using DESeq2, but the gene was highly significant when using the FL regression, as shown in Table 2.14. *SLC1A6*, which is highly expressed in the cerebellum of human brain compared to other brain regions(Furuta et al. 1997), showed lower levels of expression in prior studies of mood disorder diseases such as bipolar and major depression disorders in the striatum in situ hybridization study (McCullumsmith 2002). Furthermore, the *SLC1A6* is a member of glutamate transporter where one of the members (*SLC1A2*) showed significantly low expression in the striatum of HD samples in situ hybridization study (Arzberger et al. 1997). In addition, Utal et al. showed that Purkinje cell protein 4(*PCP4*), also known as *PEP-19*, had dramatic reduction in HD(Utal et al. 1998). This gene was not significantly associated with HD status when using DESeq2 (p-value = 0.086) but showed strong association when using FL regression (p-value = 1.15 x 10^{-4}).

Furthermore, we found that some highly expressed genes in both cases and controls may not be detected in the NB framework, because the NB framework utilizes the ratio of mean expressions of cases and controls. For instance, the normalized mean expression value of *SPOCK2* is 12,370 in cases and 15,649 in controls. Although the difference of the means is very large, the gene might not be statistically significant due to the small effect size ($\log_2$ fold-change = -0.34) in the NB framework. However, this gene is strongly associated with HD in our logistic framework as shown in Table 2.14. It is reported that the *SPOCK2* gene expression levels were significantly down regulated in high-grade astrocytoma samples.(MacDonald et al. 2007)

The top genes that showed associations exclusively in NB GLM, except for gene *AC079959.1*, have low average counts as shown in Table 2.15. The estimated dispersions for these genes are also fairly large ($\hat{\phi} > 0.5$). These genes require further investigations to be called true DE genes.

These results showed that some differently expressed genes may not be identified in the NB framework but are able to show statistical significances in the logistic framework. Moreover, the large p-values of some genes in the logistic framework impugns statistical evidence of association in the NB framework.

We recommend implementing the DA method as part of the analysis of RNA-Seq data to appropriately control Type-I error rates. If computational burden of permutations required for the DA method precludes using this approach, the FL regression is the best option for controlling Type-I errors with comparable power.

Chapter 3 Evaluation of Effect of Covariates for Case-Control Study in RNA-Seq Analysis

3.1 Introduction

An important component of differential expression analysis is to adjust for confounders. Adjustment for confounders is crucial in protecting against spurious associations. We define a confounder as a covariate that is associated with both experimental and explanatory variables. Covariates used in RNA-Seq analysis are associated with disease status, technical artifacts from experiments, or intrinsic biological properties of RNA-Seq models. If these covariates affect the abundance measurements of gene expression, then they consequently could significantly confound the association between RNA-Seq and disease status.

In the prior chapters, we considered two approaches for differential expression analysis: 1) Negative Binomial (NB) regression where gene expression is the outcome variable and case-control status is the predictor variable and 2) logistic regression where case-control status is a function of gene expression. First, we discuss covariates in the NB setting. If covariates associated with a disease status also are associated with gene expression, these covariates are confounders. However, if disease-associated covariates are not associated with gene expression, then these covariates are non-predictive (NP) covariates in models with gene expression as the outcome. Covariates that are not associated with the dependent variable (gene expression) but are associated with the independent variable (disease status) in the NB model are defined as NP covariates. Adjusting for covariates when the relationship with gene expression is unknown has not been extensively evaluated in RNA-Seq studies using the NB framework. If we alternatively consider a logistic model, the NP covariates in the NB model become non-confounding predictive (NCP) covariates in the logistic model, because the covariates are not associated with the independent variable (gene expression) but are associated with the dependent variable (disease status).

The effect of including covariates has been previously described in the Classical Logistic (CL) regression setting in the context of Genome-wide association studies (GWAS) (L. D. Robinson and Jewell 1991; Mefford and Witte 2012; Pirinen, Donnelly, and Spencer 2012) but have not been explored in the context of differential expression studies. Simulation and a real data set are used to assess the effect of including different types of covariates (NP and NCP) in NB or logistic models.

3.2 Analysis methods for evaluating effect of covariates

In our simulation, we used the NB regression model that is described in Section 2.3.1. This model includes a dispersion parameter. We utilized maximum likelihood and quasi-likelihood approaches for estimating dispersion parameters as described in Section 1.2 and Section 2.2. Additionally, we used the true dispersion parameter as set in the simulation. In our real data application, we conducted the analysis with DESeq2 that implements a NB generalized linear model detailed in Section 1.2. To compare with the NB framework, we also applied Firth's logistic (FL) regression (Section 2.3.4) was also applied to both simulated and real data sets. For both the NB and logistic models, we also implemented the data adaptive (DA) method described in Section 2.4 while analyzing the simulated data sets in order to obtain a recalibrated distribution of test statistics. The asymptotic distribution of test statistics may not be suitable for analyses when the sample size is small.

3.3 Simulation study

The simulation scenarios considered important aspects of RNA-Seq data as well as covariates. The simulation design varied sample size, mean expression value (μ), $\log_2$ fold-change (l2fc), dispersion, covariate-case status odds (CovOR), and the number of NP/NCP covariates in a model. The parameter values are provided in Table 3.1. We simulated 10,000 replicates per scenario.

Table 3.1 Parameters and their values in simulation scenarios

Parameter	**Values**
Design	Balanced, Unbalanced2, Unbalanced4
Number of cases ($N_{D=1}$)	10, 25, 75, 500
Mean expression value in controls($\mu_{D=0}$)	50, 100, 1000, 10000
Dispersion	0.01, 0.01, 0.5, 1
Covariate OR	1, 1.2, 3, 5, 10

$\log_2$ fold-change (l2fc)	0, 0.3, 0.6, 1.2, 2
Number of Covariates	0, 1, 2, 3, 5, 10

Design: Balanced has the same number of cases and controls. Unbalanced2 (4) has the 2 (or 4) times more number of controls than number of cases. Covariate OR: The odds ratio between covariates and case-control status. $\log_2$ fold-change: The l2fc equals to $\log_2\left(\frac{\text{mean expression value in cases } (\mu_{D=1})}{\text{mean expression value in controls } (\mu_{D=0})}\right)$

3.3.1 Generation of simulated RNA-Seq data

The same procedure described in Section 2.5.1 was followed to generate simulated RNA-Seq data.

3.3.2 Generation of simulated covariate data

The covariates (***X***) were simulated to follow a binomial distribution conditioning on a case-control status of subjects. The conditional probability was calculated based on the CovOR.

$$\boldsymbol{X}|D \sim B(N_D, P_D),$$

where *D* is disease status (control is 0; case is 1), N_D is sample size of *D*, $P_{D=0} = 0.5$, and $P_{D=1} = \mathrm{CovOR}/(\mathrm{CovOR} + 1)$. A set of covariates was generated based on the pre-specified CovOR. Then, using this covariate data set, additional covariates were included in a model. Of 10,000 replications in each scenario, every 10 replications were analyzed with a newly generated covariate set to incorporate within and between variances of covariates. In total, 1000 simulated covariate sets were generated per scenario. All covariates in a model were independent from each other and had the same CovOR.

3.3.3 Analysis of simulated RNA-Seq data with simulated covariates

For the NB regression we considered three different dispersion parameters. We analyzed the data using the maximum-likelihood and quasi-likelihood dispersion estimates and the true value used in the simulations. Models 3.A and 3.B define the NB and the FL regression models, respectively.

$$\text{Model 3.A: } \log(\mathrm{E}[Y]) = \beta_0 + \beta_1 D + (\textstyle\sum_{k=1}^{C} \beta_{k+1} X_k),$$

$$\text{Model 3.B: } \operatorname{logit}(\mathrm{E}[D]) = \beta_0^* + \beta_1^* Y + (\textstyle\sum_{k=1}^{C} \beta_{k+1}^* X_k),$$

where *Y* is gene expression values, *D* is a case-control status, *X* is a covariate, and *C* is the number of covariates in a model.

Type-I error rates within each scenario were calculated using the equations (2.1) at significance (alpha) levels 0.05 and 0.01. Considering the different Type-I error rates observed between NB and FL regressions, an empirical power shown in the equation (2.2) was computed with empirical threshold defined in the equation (2.3) that was calculated based on the observed Type-I error rates.

3.3.4 Cross-validation of data adaptive method in simulated RNA-Seq data

Because the expected asymptotic distribution of test statistics could not be achieved when sample size is small, we used the DA method to generate a re-calibrated distribution of test statistic based on permutations. We applied the cross-validation technique to calculate Type-I error rates. A detailed description of the cross-validation technique in the simulation study is presented in Section 2.5.3.

3.4 Simulation result

3.4.1 Type-I error simulation result

To evaluate the effect of inclusion of covariates in the models, we present Type-I error rates from the simulated data in Tables 3.2 - 3.5. As shown in Table 3.2, Type-I error rates with distinct dispersions are almost identical at both significance levels. When sample size is small (Table 3.3), an increasing number of NP covariates do not increase Type-I error rates in the NB models. The number of covariates appears to increase Type-I error rates when dispersion is 0.01 and CovOR is 5 (Table 3.3). However, this slightly increased Type-I error rate is close to the Type-I error rate without any covariates in the model when dispersion is 0.01 and CovOR is 1.2. Adding more NP covariates when the dispersion is large increases Type-I error rates. However, the effects of large CovOR on Type-I error rates are not notable in NB models. Large sample size (Table 3.3 and Table 3.4) weakens the inflation that arises from a large number of NP covariates within large dispersion in the NB model.

As defined in Chapter 3.1, the same covariates that are NP covariates in an NB model are NCP covariates in logistic models. Unlike the NB regression, even with small sample size (Table 3.2), when the CovOR is small, the FL regression is robust with the increment of the number of NCP covariates. When CovOR is large, Type-I error rates from FL regression become very conservative as the number of NCP covariates increase. Type-I error rates are not affected by large dispersion. When sample size increases, Type-I error rates at both significant levels are less affected by increased number of NCP covariates with large CovOR (Tables 3.3 and 3.4).

Table 3.2 Type-I error rates of the NB regressions with the true dispersion and ML and QL dispersions from balanced design of 10 cases and 1000 mean expressions

			$\alpha = 0.05$			$\alpha = 0.01$		
Disp	**CovOR**	**Ncov**	**NB_TD**	**NB_MLD**	**NB_QLD**	**NB_TD**	**NB_MLD**	**NB_QLD**
0.01	1	0	0.066	0.066	0.066	0.017	0.017	0.017
0.01	1	3	0.063	0.063	0.063	0.018	0.018	0.018
0.01	1	5	0.067	0.067	0.067	0.021	0.021	0.021
0.01	5	0	0.064	0.064	0.064	0.016	0.016	0.016
0.01	5	3	0.073	0.073	0.073	0.021	0.020	0.020
0.01	5	5	0.080	0.080	0.080	0.024	0.024	0.024
1	1	0	0.090	0.090	0.090	0.030	0.030	0.030
1	1	3	0.128	0.128	0.128	0.050	0.050	0.051
1	1	5	0.152	0.152	0.152	0.066	0.066	0.066
1	5	0	0.088	0.088	0.088	0.030	0.030	0.030
1	5	3	0.127	0.127	0.127	0.050	0.050	0.050
1	5	5	0.142	0.142	0.142	0.061	0.061	0.061

Disp: Dispersion, CovOR: Odds ratios between covariates and case-control status, Ncov: The number of covariates in a model, NB: Negative Binomial, TD: The dispersion is used for the sampling, MLD: Maximum likelihood estimated Dispersion, QLD: Quasi-likelihood estimated Dispersion

Table 3.3 Type-I error rates of the NB and Firth's logistic regressions from balanced design of 10 cases and 1000 mean expressions

			$\alpha = 0.05$		$\alpha = 0.01$	
Disp	**CovOR**	**Ncov**	**NB_TD**	**FL**	**NB_TD**	**FL**
0.01	1	0	0.066	0.045	0.017	0.007
0.01	1	3	0.063	0.043	0.018	0.008
0.01	1	5	0.067	0.044	0.021	0.008
0.01	1.2	0	0.072	0.048	0.022	0.009
0.01	1.2	3	0.072	0.049	0.024	0.010
0.01	1.2	5	0.076	0.050	0.026	0.009
0.01	5	0	0.064	0.042	0.016	0.008
0.01	5	3	0.073	0.036	0.021	0.004
0.01	5	5	0.080	0.021	0.024	0.001
1	1	0	0.090	0.040	0.030	0.006
1	1	3	0.128	0.043	0.050	0.007
1	1	5	0.152	0.045	0.066	0.008
1	1.2	0	0.091	0.040	0.031	0.007
1	1.2	3	0.128	0.043	0.054	0.007
1	1.2	5	0.151	0.045	0.067	0.008
1	5	0	0.088	0.038	0.030	0.007
1	5	3	0.127	0.034	0.050	0.005
1	5	5	0.142	0.020	0.061	0.001

Disp: Dispersion, CovOR: Odds ratios between covariates and case-control status, Ncov: The number of covariates in a model, NB_TD: The Negative binomial regression with the dispersion used for the sampling, FL: Firth's Logistic regression

Table 3.4 Type-I error rates of the NB and Firth's logistic regressions from balanced design of 25 cases and 1000 mean expression values

			$\alpha = 0.05$		$\alpha = 0.01$	
Disp	**CovOR**	**Ncov**	**NB_TD**	**FL**	**NB_TD**	**FL**
0.01	1	0	0.054	0.046	0.014	0.009
0.01	1	3	0.057	0.048	0.014	0.010
0.01	1	10	0.058	0.052	0.014	0.010
0.01	1.2	0	0.057	0.048	0.013	0.009
0.01	1.2	3	0.057	0.049	0.014	0.010
0.01	1.2	10	0.058	0.052	0.014	0.010
0.01	10	0	0.052	0.045	0.010	0.006
0.01	10	3	0.058	0.046	0.013	0.008
0.01	10	10	0.052	0.001	0.014	<0.001
1	1	0	0.065	0.043	0.017	0.007
1	1	3	0.080	0.044	0.022	0.007
1	1	10	0.116	0.050	0.037	0.008
1	1.2	0	0.065	0.044	0.019	0.009
1	1.2	3	0.079	0.046	0.025	0.009
1	1.2	10	0.116	0.049	0.041	0.010
1	10	0	0.066	0.046	0.019	0.008
1	10	3	0.082	0.040	0.023	0.007
1	10	10	0.122	0.004	0.045	<0.001

Disp: Dispersion, CovOR: Odds ratios between covariates and case-control status, Ncov: The number of covariates in a model, NB_TD: The Negative binomial regression with the dispersion used for the sampling, FL: Firth's Logistic regression

Table 3.5 Type-I error rates of the NB and Firth's logistic regressions from balanced design of 75 cases and 1000 mean expression values

			$\alpha = 0.05$		$\alpha = 0.01$	
Disp	**CovOR**	**Ncov**	**NB_TD**	**FL**	**NB_TD**	**FL**
0.01	1	0	0.055	0.052	0.013	0.011
0.01	1	3	0.055	0.052	0.013	0.011
0.01	1	10	0.054	0.052	0.012	0.011
0.01	1.2	0	0.053	0.050	0.013	0.012
0.01	1.2	3	0.053	0.049	0.014	0.013
0.01	1.2	10	0.053	0.051	0.013	0.012
0.01	10	0	0.052	0.049	0.012	0.010
0.01	10	3	0.051	0.045	0.011	0.009
0.01	10	10	0.052	0.035	0.011	0.004
1	1	0	0.057	0.050	0.013	0.009
1	1	3	0.062	0.049	0.015	0.010
1	1	10	0.073	0.050	0.019	0.010
1	1.2	0	0.054	0.046	0.013	0.010
1	1.2	3	0.058	0.046	0.015	0.010
1	1.2	10	0.067	0.049	0.020	0.011
1	10	0	0.057	0.048	0.013	0.009
1	10	3	0.063	0.049	0.016	0.010
1	10	10	0.077	0.038	0.022	0.008

Disp: Dispersion, CovOR: Odds ratios between covariates and case-control status, Ncov: The number of covariates in a model, NB_TD: The Negative binomial regression with the dispersion used for the sampling, FL: Firth's Logistic regression

3.4.2 Type-I error simulation result with DA method

In all scenarios, the DA method controls Type-I error rates well in the NB and FL regressions at both 0.05 and 0.01 alpha levels as presented in Table 3.6. The newly approximated distribution of test statistics diminishes deviated Type-I error rates that were not controlled when many covariates were included in the NB and FL models.

Table 3.6 Type-I error rates of the NB and Firth's logistic regressions with DA method from balanced design of 1000 mean expressions

				$\alpha = 0.05$		$\alpha = 0.05$	
Ncase	**Disp**	**CovOR**	**Ncov**	**NB_TD**	**FL**	**NB_TD**	**FL**
10	1	5	5	0.042	0.052	0.011	0.010
25	1	10	10	0.047	0.043	0.010	0.011
75	1	10	10	0.046	0.049	0.010	0.011

Ncase: The number of cases; and the number of controls are the same, Disp: Dispersion, CovOR: Odds ratios between covariates and case-control status, Ncov: The number of covariates in a model, NB_TD: The Negative binomial regression with the dispersion used for the sampling, FL: Firth's Logistic regression

3.4.3 Empirical power simulation result

The results of the power simulations to evaluate the inclusion of covariates in the models are summarized in Figures 3.1 - 3.3. Similar to the Type-I error rate, the empirical power of the NB regressions using different dispersion estimation methods were similar for all power scenarios. When sample size is increased the overall power is increased (Figure 3.1 - 3.3) in both NB and FL regression.

In our simulation, when sample size is fixed, the power of NB and FL regression is affected by three factors 1) Dispersion, 2) CovOR, and 3) The number of NP/NCP covariates in a model. Large dispersion, large CovOR, and increasing number of NP/NCP covariates in a model decrease power. Power of NB regression is less sensitive to the increase of NP covariates with small dispersion than with large dispersion. As shown in Figure 3.1(A), NB regression shows marginally more power than FL regression when the number of covariates is large. When dispersion is large but CovOR is small, the loss of power in NB regression is more sensitive to the increase of the number of NP covariates than in FL regression as seen in Figure 3.1(D) and Figure 3.2(D). In particular, with CovORs of 1 or 1.2, the power of FL regression with 10 covariates in a model is more powerful than NB regression with 10 covariates. Regardless of dispersion, when CovOR and the number of covariates in a model are large, NB regression shows better power than FL regression. This is demonstrated in Figure 3.1 with CovOR equal to 5 and Figure 3.2 with CovOR equal to 10.

Figure 3.1 Empirical power of NB and FL regressions with covariates for a balanced design with 10 cases and mean expression in controls of 1000

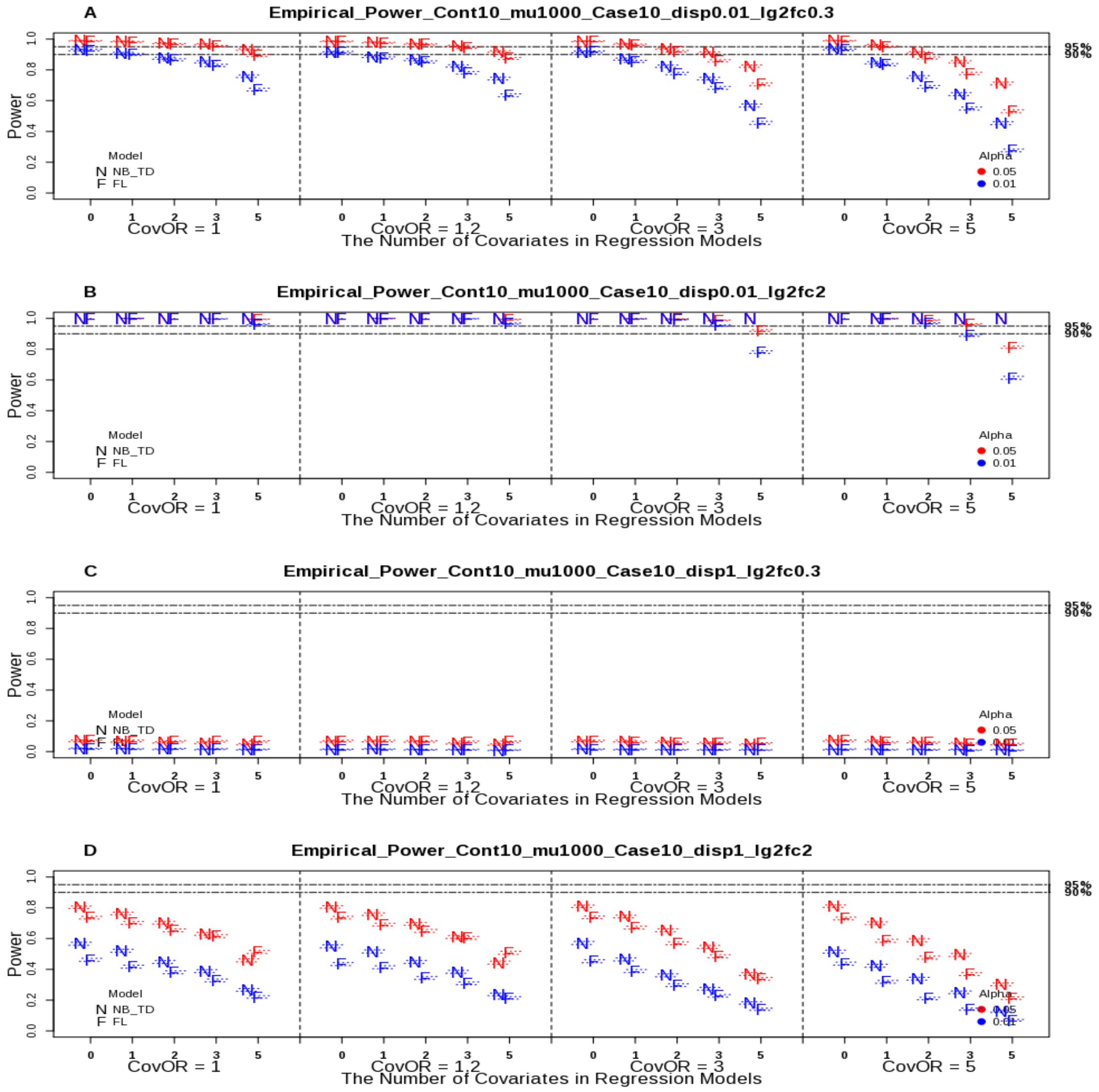


Figure 3.1 contains power of the Negative Binomial with true dispersion (NB_TD) and Firth's Logistic (FL) regressions at alpha levels of 0.05 and 0.01. The black dotted horizontal lines represent 95% and 90% of power from the top. The odds ratios between covariates and case-control status (CovOR = 1, 1.2, 3, and 5) are separated by black dotted vertical lines. Five values of the number covariates in a model (0, 1, 2, 3, and 5) are placed within each mean expression value. Dotted lines within each character imply 95% confidence interval of p.value. (A) and (B) have l2fc values of 0.3 and 2 within the same dispersion of 0.01. (C) and (D) have l2fc values of 0.3 and 2 within the same dispersion of 1.

Figure.3.2 Empirical power of NB and FL regressions with covariates for a balanced design with 25 cases and mean expression in controls of 1000

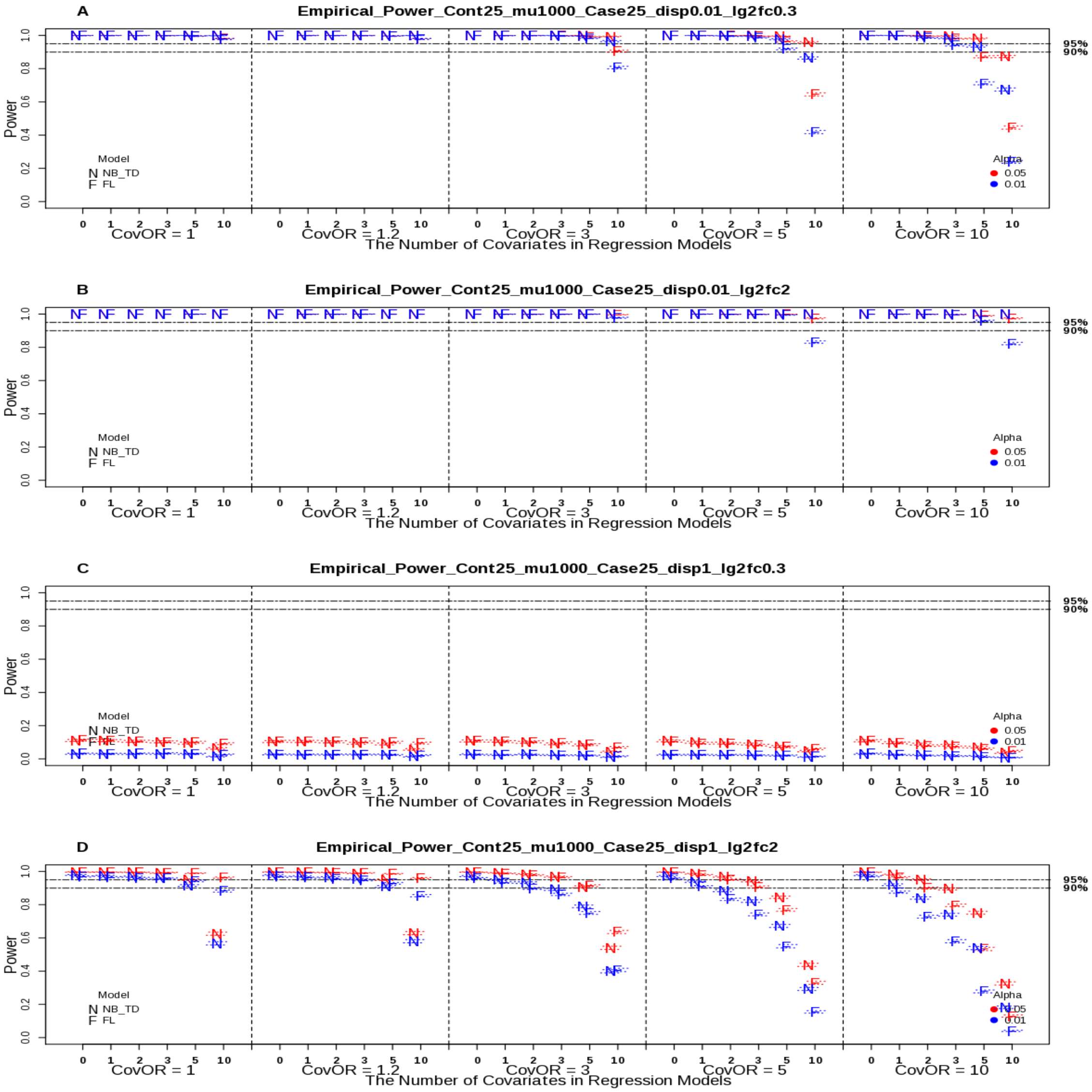


Figure 3.2 contains power of the Negative Binomial with true dispersion (NB_TD) and Firth's Logistic (FL) regressions at alpha levels of 0.05 and 0.01. The black dotted horizontal lines represent 95% and 90% of power from the top. The odds ratios between covariates and case-control status (CovOR = 1, 1.2, 3, 5, and 10) are separated by black dotted vertical lines. Six values of the number covariates in a model (0, 1, 2, 3, 5, and 10) are placed within each mean expression value. Dotted lines within each character imply 95% confidence interval of p.value. (A) and (B) have l2fc values of 0.3 and 2 within the same dispersion of 0.01. (C) and (D) have l2fc values of 0.3 and 2 within the same dispersion of 1.

Figure 3.3 Empirical power of NB and FL regressions with covariates for a balanced design with 75 cases and mean expression in controls of 1000

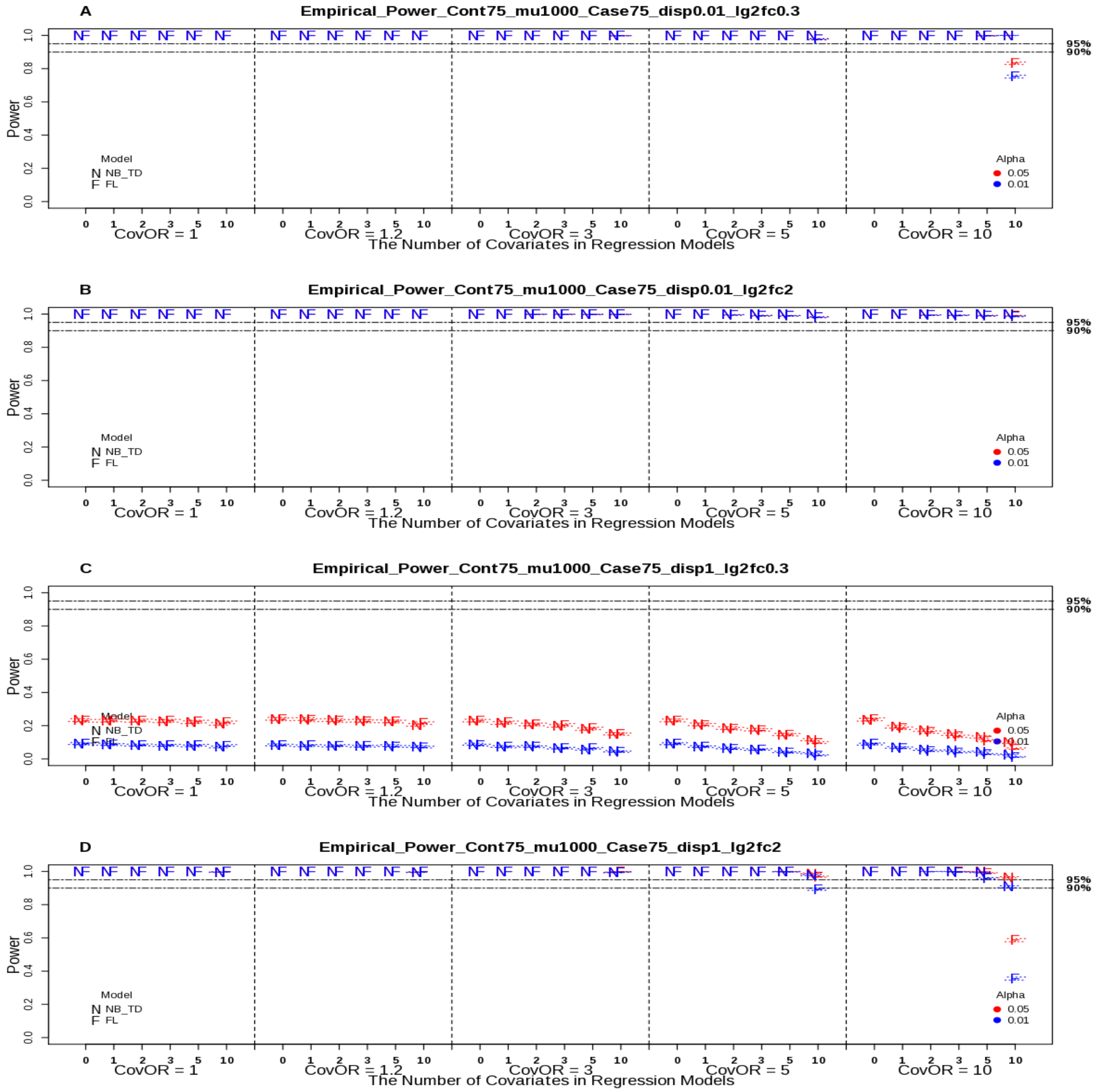


Figure 3.3 contains power of the Negative Binomial with true dispersion (NB_TD) and Firth's Logistic (FL) regressions at alpha levels of 0.05 and 0.01. The black dotted horizontal lines represent 95% and 90% of power from the top. The odds ratios between covariates and case-control status (CovOR = 1, 1.2, 3, 5, and 10) are separated by black dotted vertical lines. Six values of the number covariates in a model (0, 1, 2, 3, 5, and 10) are placed within each mean expression value. Dotted lines within each character imply 95% confidence interval of p.value. (A) and (B) have l2fc values of 0.3 and 2 within the same dispersion of 0.01. (C) and (D) have l2fc values of 0.3 and 2 within the same dispersion of 1.

3.5 Application to the real RNA-Seq data set of Huntington's Disease (HD)

Details of the HD data set that has 20 cases and 49 controls are described in Section 2.6

3.5.1 Analysis of HD RNA-Seq data with simulated covariates

To evaluate the effect of covariates in a model, the same method for generating covariates in our simulation study was applied to the HD data set to create simulated covariates. In this real data application, we focused on a moderate and realistic covariate effect on HD status (CovOR = 1.2)

The original HD data with simulated covariates were analyzed using the NB generalized linear model in DESeq2 with Model 3.C and using the FL regression with Model 3.D.

$$\text{Model 3.C: } \log_2(\mathrm{E}[Y]) = \beta_0 + \beta_1 D + \beta_2 ADD_{1\text{vs}.2} + \beta_3 ADD_{1\text{vs}.3} + \beta_4 ADD_{1\text{vs}.4} + \beta_5 RIN + \left(\sum_{k=6}^{C+5} \beta_k X_k\right),$$

$$\text{Model 3.D: } \mathrm{logit}(\mathrm{E}[D]) = \beta_0^* + \beta_1^* Y + \beta_2^* ADD_{1\text{vs}.2} + \beta_3^* ADD_{1\text{vs}.3} + \beta_4^* ADD_{1\text{vs}.4} + \beta_5^* RIN + \left(\sum_{k=6}^{C+5} \beta_k^* X_k\right),$$

where *C* is the number of simulated covariates, and *C* = 1, 2, 3, 5, or, 10. The change of genomic inflation factors with the addition of a varying number of simulated covariates in a model was evaluated.

3.5.2 Result of HD RNA-Seq data with simulated covariates

The HD data was analyzed using DESeq2 with additional NP covariate models and using FL regression with additional NCP covariate models. The summary of genomic lambdas is presented in Table 3.7. An increase of NP/NCP covariates leads to a marginally lower genomic inflation factor. The standard deviations of the genomic inflation factors are increased with the increase of NP/NCP covariates.

Table 3.7 Summary of genomic inflation factor from HD analyses with simulated covariates

Method	Ncov	Median	SD
NB	1	4.046	0.095
	2	4.021	0.139
	3	3.998	0.166
	5	3.931	0.215
	10	3.744	0.293
FL	1	3.504	0.155
	2	3.463	0.222
	3	3.404	0.273
	5	3.281	0.352
	10	2.949	0.525

NB: Negative binomial regression implemented in DESeq2, FL: Firth’s logistic regression, Ncov: The number of simulated covariates in the model, SD: Standard deviation

3.6 Discussion

The effect of NCP covariates was investigated in the context of GWAS by Pirinen et al.(Pirinen, Donnelly, and Spencer 2012) using classical logistic regression. They demonstrated that NCP covariates that are known to be associated with a disease outcome may reduce power to identify associations between the disease and genetic variants. Later, an improved method using the liability threshold model with an informed relationship between disease and covariates was proposed by Zaitlen et al.(Zaitlen et al. 2012) in GWAS with a case-control study design, but the effect of including covariates has not been investigated for RNA-Seq studies. The statistical relationship between covariates and disease status could be conveniently identified through an individual association test or a multivariate association test. However, identifying relationships with covariates for all genes is computationally demanding. Existing software does not allow defining gene-wise models for all genes, which makes this approach challenging for many researchers. Therefore, RNA-Seq studies that include covariates in a single model applied to all genes will likely result in some gene expression models that include unassociated covariates. Hence, it is important to investigate the effect of NP covariates for gene expression in RNA-Seq analysis.

Simulations that included NP covariates in the NB model showed inflated Type-I error rates and a loss of power. With large dispersion, this inflation and loss of power becomes severe.

The Type-I error in the FL regression is not notably affected by the increment of the number of NCP covariates when CovOR is small. With large CovOR and increased number of NCP covariates, conservative Type-I error rates are observed. The DA method effectively controls the increase of Type-I error rates even with larger CovOR and high number of NP/NCP covariates. Our analysis of empirical power shows that the FL regression is more greatly influenced by the increase of covariates than the NB regression, when CovOR is large.

Our HD analyses with simulated NP/NCP covariates demonstrated that an increase in the number of NP/NCP covariates results in the increased variability of the genomic inflation factor (Table 3.7). Adding more NP covariates to an NB model slightly decreases the median of the genomic inflation factor. The decreased genomic inflation factor indicates the increased median of p-values. In other words, many p-values in a set are generally increased. In our simulation, we found large dispersion significantly increases Type-I error rates, as the number of NP covariates in a model increases. Also, power is significantly decreased as the number of NP covariates in a model increases. The increased Type-I error rates imply decreased p-values, and the decreased power indicate increased p-values. Therefore, this slightly decreased median of the genomic inflation factor may indicate that the loss of power is greater than the gain of Type-I error rates.

Adding more NCP covariates in a model also slightly decreases the median of the genomic inflation factor in FL regression. This decreased median might be caused by the loss of power, which may result from NCP covariates in a model according to our simulation results. Under a moderate CovOR, the number of NCP covariates in a model does not affect the Type-I error rates.

The change in the median of the genomic inflation factor with additional covariates is larger in FL regression than NB regression because the FL regression results are solely affected by the loss of power (Table 3.7). NB regression results are influenced by both increased Type-I error and decreased power.

The standard deviation of genomic inflation factor is increased with adding NP/NCP covariates in a model. This means that the results generated from a model that includes many covariates is unreliable, even if these covariates are associated with case-control status but not gene expression.

When covariates are not significantly associated with the expression of a particular gene, their inclusion in the model may cause spurious association, and miss true differential expression. Although it is not ideal to design a separate model for each gene by identifying association between the gene expression and covariates, we need to be cautious that the unknown relationship between expression level of a gene and covariates may induce a false association.

In conclusion, adding disease-associated covariates to a model may not control Type-I error rates or improve power. Although the DA method is able to control Type-I error rates, the computational burden of performing permutations for each gene may prevent researchers from utilizing the DA method. However, if the covariates in a model do not have strong relationship with case-control status, Type-I error rates can be controlled in FL regression. This is in contrast to the NB approach where the effect of dispersion on Type-I error rates cannot be controlled. The loss of power cannot be avoided in both NB and FL regressions if included covariates are NP or NCP covariates. Therefore, a parsimonious model with FL regression is recommended in RNA-Seq studies.

Chapter 4 Multiple testing correction methods in RNA-sequence data

4.1 Introduction

Multiple testing corrections are an important procedure when many hypotheses are tested across an entire set of high dimensional data such as genetic or genomic data. To control Type-I error, different approaches, that control Type-I error in distinct ways, may be utilized depending on the study. Some multiple testing correction approaches control the familywise error rate (FWER) while others control the false discover rate (FDR). The FWER is defined as the probability of one or more Type-I errors in a family of hypothesis tests, and the FDR can be defined as the expected proportion of errors among the rejected hypotheses(Benjamini and Hochberg 1995). Benjamini and Hochberg (BH) proposed a procedure for controlling the FDR(Benjamini and Hochberg 1995), and application of this procedure or one of the modifications to the BH approach (Storey and Tibshirani 2003; Benjamini, Krieger, and Yekutieli 2006) is a common strategy to address multiple testing issues in differential expression studies. It is known that many genes are co-expressed and hence, their expression values have a complex dependence structure(Stuart 2003). These regulatory processes involve multiple genes and together create a complex regulatory network. Jain et al. showed that a Bonferroni procedure is very conservative to control FWER in microarray data(Jain et al. 2003). The reason is that whereas this procedure is most robust for independent tests, the expression levels among many genes are correlated. Hence, the correlation among genes should to be taken into account to control the FWER.

The BH procedure for controlling FDR successfully provides a reasonable balance between true and false positives when applied to microarray data. Microarray technology measures the fluorescence of targeted RNA molecules. These processed measurements generally follow a normal distribution. However, RNA-sequencing (RNA-Seq) reads generated from next generation sequencing technologies are becoming more widely used because this technology provides counts of targeted RNA molecules. Consequently, appropriate statistical methods have to be developed to analyze RNA-Seq data. DESeq2 and edgeR are two popular R-packages(Love, Huber, and Anders 2014; M. D. Robinson, McCarthy, and Smyth 2010) that adopt the Negative Binomial (NB) framework to analyze RNA-Seq count data. As an alternative to the NB approach, Firth’s logistic regression has been described in Chapter 2 and Chapter 3.

Following analysis of this high dimensional RNA-Seq data, the BH procedure has been widely used as a method for adjusting for multiple testing in RNA-Seq analysis and this procedure was implemented in DESeq2 and edgeR as a default option. However, the dependence structures in transcriptomic data may violate the assumptions in the BH procedure, which requires the positive regression dependent on a subset condition. Correlations among RNA-seq count data genes pairs can be both positive and negative.

Since the publication of the BH procedure, researchers have proposed refinements. One area of interest is to estimate the proportion of null hypotheses in multiple testing inferences. Incorporating this estimated null proportion in the procedure to control FDR has been shown to be more powerful than the BH method(Storey 2002; Black 2004). Storey and Tibshirani (2003), Nettleton et al (2004), and Pounds and Cheng (2006) proposed methods to estimate the proportion of null hypotheses(Storey and Tibshirani 2003; Nettleton et al. 2006; Pounds and Cheng 2006). Dialsingh et al. evaluated the performance of methods estimating the proportion of null hypotheses using RNA-Seq data (Dialsingh, Austin, and Altman 2015).

Multiple studies have explored issues related to FDR in the analysis of RNA-Seq (Burden, Qureshi, and Wilson 2014; Dialsingh, Austin, and Altman 2015; Rocke et al. 2015; Li et al. 2012; Si and Liu 2013). Burden et al. showed that p-values from null hypotheses frequently do not follow a uniform distribution in over-dispersed RNA-Seq data. The non-uninform distribution of p-values from null hypotheses leads to inaccurate FDR estimates(Burden, Qureshi, and Wilson 2014). Rocke et al. favor the use of a critical level (1 x 10^{-4}) for multiple comparisons over FDR procedures because the FDR adjusted p-values are a complex function of an entire vector of p-values and these adjusted p-values are difficult to interpret (Rocke et al. 2015). Li et al. and Si et al. proposed novel statistical inference tests to identify differentially expressed genes. They also proposed improved procedures for estimating FDRs in RNA-Seq studies(Li et al. 2012; Si and Liu 2013). These studies strongly suggest that inappropriate usage of multiple testing correction methods that control FDR may lead to spurious conclusions.

Lehmann and Romano(Lehmann and Romano 2005) and Romano and Shaikh(Romano and Shaikh 2006) proposed methods controlling FWER which were free of an independence assumption while maintaining reasonable power. Gao et al. also provided a method to compute the effective number of independent tests and adjust the FWER in the context of GWAS(Gao, Starmer, and Martin 2008). This method has been shown to be more powerful than other methods that estimate the effective number of tests(Hendricks et al. 2014). However, this method cannot be implemented when the sample size is small, which is common in many RNA-Seq studies, including our HD example in Chapters 2 and 3. Small sample size impedes the correct estimation of eigenvalues that are required to determine the effective number of tests. If the sample size is large (500 or more ,as shown in an example data in *http://simplem.sourceforge.net/*), estimating the effective number of tests can be an attractive method.

Many methods exist to control FDR for data sets with dependence structures. Benjamini and Yekutieli(Yekutieli and Benjamini 2001) provided an FDR correction method that accounts for dependence structures. Storey and Tibshirani(Storey and Tibshirani 2003) also showed their method is powerful in the presence of weak correlation structures. Benjamini, Krieger, and Yekutieli(Benjamini, Krieger, and Yekutieli 2006) showed their method performed well in data sets with a dependence structure. Blanchard and Roquain(Blanchard and Roquain 2008) proposed a method applicable for any type of dependence structure. Each of these methods is described in greater detail below (Sections 4.2.1 and 4.2.2).

Although many multiple testing correction methods that are applicable to correlated data have been developed to control FWER and FDR, these methods have not been exhaustively investigated in the context of RNA-Seq data. In particular, scenarios in which correlation exists among genes that are not differentially expressed (under the null hypothesis) have not been investigated. In this chapter, we compare the performance of multiple testing correction methods using simulated RNA-Seq data sets containing correlated gene expression measures.

4.2 Methods

The multiple testing correction methods that control FWER are evaluated in terms of false positive rates and power (1 - false negative rates). The multiple testing correction methods that impose FDR are evaluated in terms of false discovery rates and power.

4.2.1 Multiple testing correction procedures controlling FWER

We used the Bonferroni procedure as the reference method that controls the FWER(Bonferroni 1936). The critical value ($\alpha^{Bonferroni}$) of Bonferroni procedure is compared with p-values, where

$$\alpha^{Bonferroni} = \frac{\alpha}{m} ,$$

α is a nominal significance level, and *m* is the number of tests.

4.2.1.1 Lehmann and Romano (LR) procedure

k-FWER has been defined as the probability of having *k* or more false positives(Lehmann and Romano 2005).

$$k - FWER = \Pr\{\text{ reject at least k hypotheses } \mathrm{H}_i \text{ with } i \ \in \mathrm{I(P)}\}$$

where I(P) is the set of true null hypotheses when P is the true probability distribution. Control of the k-FWER requires that k-FWER < α for all P. The LR method provides a generalized step-down procedure and controls the FWER under any dependence structure. The *j*th ordered p-value among *m* individual tests is compared with the step-down constant α_j^{LR}, where

$$\alpha_j^{LR} = \begin{cases} \frac{k\alpha}{m} & if\ j \leq k \\ \frac{k\alpha}{m+k-1} & if\ j > k \end{cases}.$$

4.2.1.2 Romano and Shaikh (RS) procedure

The RS approach proposes a generalized step-up procedure (Romano and Shaikh 2006). Like the LR procedure, the RS procedure controls the FWER under all dependence conditions. The RS procedure compares the j^{th} ordered p-value with a critical value α_j^{RS}, where

$$\alpha_j^{RS} = \frac{\alpha_j^{LR}}{D_1(k,m)} \ ,$$

$$D_1(k,m) = \begin{matrix} \max \\ k < |\mathrm{I}| < m \end{matrix} \left[|I| \frac{\alpha_{m-|\mathrm{I}|-k}}{k} + |I| \sum_{k<j<|I|} \frac{\alpha_{m-|\mathrm{I}|-j} - \alpha_{m-|\mathrm{I}|-1}}{j} \right],$$

and $|\mathrm{I}|$ is the number of alternative hypotheses. Romano and Shaikh (2006) demonstrated that the critical value α_j^{RS} is approximately one-half of α_j^{LR}. This indicates that RS procedure is more conservative than the LR method, but it is not clear which is more appropriate for controlling the FWER under different dependence structures in the data.

4.2.2 Multiple testing correction procedures controlling FDR

The BH procedure is used as a reference method for controlling the FDR(Benjamini and Hochberg 1995). The critical value (α_j^{BH}) in the BH procedure is compared with the j^{th} ordered p-value, where

$$\alpha_j^{BH} = \frac{j\alpha}{m}$$

4.2.2.1 Benjamini and Yekutieli (BY) procedure

The BY approach utilizes a step-up procedure that controls FDR under any dependence structure(Yekutieli and Benjamini 2001). This procedure compares the j^{th} ordered p-value with a constant α_j, where

$$\alpha_j^{BY} = \frac{j\alpha}{m\sum_{i=1}^{m}\frac{1}{i}} \ .$$

This procedure is more conservative than the BH procedure because the critical values of BY procedure are decreased by $\sum_{i=1}^{m}\frac{1}{i}$.

4.2.2.2 Storey and Tibshirani (ST) procedure

Storey (2002) first suggested the importance of knowing the proportion of null hypotheses(Storey 2002). P-values for true alternative hypotheses presumably are near to zero, but p-values of null hypotheses should follow a [0,1] uniform distribution. Thus, the overall proportion of null p-values can be estimated as

$$\hat{\pi}_0(\lambda) = \frac{\#\,\{p_i > \lambda; i = 1, \dots, m\}}{m(1-\lambda)},$$

where λ is a tuning parameter.

Storey and Tibshirani proposed to estimate λ by fitting a natural spine with 3 degrees of freedom to the values of $\hat{\pi}_0(\lambda)$ with λ ranging from 0.01 to 0.95(Storey and Tibshirani 2003). With the estimated $\hat{\pi}_0$, the *j*th q-value is calculated as

$$\hat{q}\left(p_{(j)}\right) = \begin{matrix} \min \\ t > p_{(j)} \end{matrix} \frac{\hat{\pi}_0 m t}{\#\{p_j \leq t\}},$$

where t is a threshold between 0 and 1. Although the ST procedure assumes independence of p-values, this procedure performed well in various simulations under dependency(Storey and Tibshirani 2003).

4.2.2.3 Benjamini, Krieger, and Yekutieli (BKY) procedure

Benjamini et al (2006) suggested a two-stage procedure to control FDR(Benjamini, Krieger, and Yekutieli 2006). This BKY procedure is an adaptive version of the BH procedure. First, the BKY procedure estimates the number of null hypotheses, m_0, with following function

$$\widehat{m}_0^{BKY} = \frac{m - r_1}{1 - q} = (m - r_1)(1 + \alpha),$$

where $q = \frac{\alpha}{1+\alpha}$, r_1 is the number of rejections, and m is the total number of tests. Then, the j^{th} p-value is compared with a critical value where

$$\alpha_j^{BKY} = \frac{j\alpha}{\widehat{m}_0^{BKY}}.$$

Although the BKY procedure was developed under the independence assumption, this procedure has shown good performance for positively dependent test statistics(Benjamini, Krieger, and Yekutieli 2006).

4.2.2.4 Blanchard and Roquain (BR) procedure

Blanchard and Roquain (2008) suggested a step-up method that controls FDR with any dependence structure(Blanchard and Roquain 2008). This procedure is a generalization of the BY procedure. The j^{th} p-value is compared with a critical value, α_j^{BR}, defined as:

$$\alpha_j^{BR} = \frac{\alpha}{m}\,\beta(j)\text{ , where}$$

$$\beta(j) \equiv \beta_v(j) = \int_0^j x\; dv(x)$$

and v is an arbitrary probability distribution on $(0, \infty)$. We used a prior distribution proportional to $\exp\,(\frac{-j}{0.15m})$ as a default prior.

4.2.3 Simulation study

We partially follow a gene set simulation method proposed by Landau and Liu (Landau and Liu 2013). This method simulates a set of gene expression measures using pairs of mean (μ_g) and dispersion ($ø_g$) of genes from a real RNA-Seq data set. This method is able to generate unstructured correlations among differentially expressed (DE) genes.

As a modification to the original simulation method, we include correlation structures among simulated null-genes. Five correlated structures are simulated as shown in Table 4.1. First, the gene sets labeled “exchangeable” have an exchangeable structure that has the same pairwise positive correlation coefficients. This exchangeable structure is often observed in clustered data. Second, a set of genes has the same absolute correlation coefficient. However, for every alternate cell by columns and rows, the sign of the coefficient is changed. We call this correlation structure “exchangeable2”. Genes within this correlation structure have 50% positive and 50% negative pairwise correlation coefficients. For the third simulated structure, we used the observed correlation structures from a real RNA-Seq data to model correlation. We call this the “real” structure. A fourth correlation structure is “autoregressive1”. When genes have an autoregressive1 structure, genes have the same variance and pair-wise correlations exponentially decrease with distance. Finally, we simulate RNA-Seq gene expression measures that have zero correlation (“independent”).

Because the assumption of dependence structures is imposed under the null hypotheses, we simulated correlation structures restricted to gene expression measured under the null hypotheses in our primary simulation. However, in our secondary simulations, we specified correlations among DE and non-DE genes.

4.2.3.1 Generation of simulated RNA-Seq data sets

We simulate whole gene sets using the following steps and the combination of parameters presented in Table 4.1.

1. Randomly select 10,000 genes from the real RNA-Seq data (Pickrell et al. 2010) without replacement. The corresponding 10,000 pairs of μ_g and $ø_g$ are used as the geometric mean expression level across treatments and true dispersion, respectively, of simulated genes.
2. Randomly select simulated genes to be either differentially expressed across the two treatments or equivalently expressed such that exactly *(π_0 x 100)*% of the simulated genes are differentially expressed and the remaining (1 - π_0) *x 100*% are equivalently expressed.
3. Set the log fold-change across treatment levels, δ_g, to be zero for all equivalently expressed genes. In order to have independent differentially expressed genes, we draw the δ_g's of all differentially expressed genes from a multivariate normal distribution with mean 0 and variance equal to an identity matrix. Although Landau and Liu suggested implementing dependent structures among DE genes, because this dependent structure does not violate the assumptions of the multiple testing correction methods, we assume independence among DE genes in our primary analysis. We generate the unstructured correlations among DE genes using the *rcorrmatrix()* function in “ClusterGeneration” R-package(Joe 2006) in our secondary analysis.
4. Compute the true mean expression level, μ_{gk}, of simulated gene, g, for treatment levels k=1 and 2 using

$$\mu_{gk} = a_g \exp\left(-1^k \frac{\delta_g}{2}\right).$$

The log fold-change can then be expressed as

$$\log(\text{fold change}) = \log\left(\frac{a_g \exp\left(\frac{\delta_g}{2}\right)}{a_g \exp\left(-\frac{\delta_g}{2}\right)}\right) = \log\left(\exp\left(\frac{\delta_g + \delta_g}{2}\right)\right) = \delta_g$$

5. Randomly draw the simulated count of each simulated gene, *g,* in library *i* from a $NB(\mu_{gk(i)}, \phi_g)$ distribution, where *k(i)* is the treatment group of library *i*.
6. Only genes with simulated read counts greater than zero are included in the following analysis. Hence, if the simulated counts of simulated gene g are all zero, we keep δ_g and redraw μ_g and $ø_g$, and then redraw the simulated counts as in steps 4 and 5.
7. We randomly select 20 genes among non-DE genes for exchangeable, exchangeable2, autoregressive1, and real correlation matrices. The selected genes simulated from the $NB(\mu_{gk(i)}, \phi_g)$ distribution and the correlation among these genes has the specified structure with correlation coefficient of ρ in Table 4.1 for exchangeable, exchangeable2, and autoregressive1. For our secondary analysis, we only use ρ equal to 0.5. The real correlation matrix is estimated from the Pickrell RNA-Seq data(Pickrell et al. 2010). Using the observed values avoids a non-positive definite correlation matrix. This sampling procedure is independently performed 150 times per replicate, to form 150 clusters of 20 genes for

each simulation replicate when the correlation structures are exchangeable, exchangeable2, autoregressive1 or real. The correlated discrete sampling method uses the *rcounts* function in the R-package “corcounts” (Vinzenz and Claudia 2009). This function allows random sampling from NB distributions with pre-specified Pearson correlation coefficient(ρ). When the correlation structure is independent, all genes are independently sampled from $NB(\mu_{gk(i)}, \phi_g)$.

Table 4.1 Simulation parameters and values

Parameter	**Value**
Sample size	5, 10, 20
Proportion of null(π_0)	1, 0.95, 0.75
Strength of correlation(ρ)	0.3, 0.5, 0.75
Correlation Structure	Exchangeable, Exchangeable2, Real, Autoregressive1, Independent

Sample size: The number of samples in case and control groups. Each group has the same sample size. Proportion of null: the proportion of null hypotheses in a whole gene set.

4.2.3.2 Analysis of simulated RNA-Seq data sets

We analyze the simulated data sets using the DESeq2 and edgeR R-packages and using Firth’s logistic regression. DESeq2 and edgeR are the leading software for analyzing RNA-Seq data. Firth’s logistic regression(Firth 1993; Heinze and Schemper 2002) is an appropriate alternative approach as demonstrated in Chapter 2 and Chapter 3 of this dissertation.

Because DESeq2 and edgeR use the NB framework, gene expression values are a function of case-control status. However, Firth's logistic regression models case-control status as a function of gene expression values. Hence, DESeq2 and edgeR utilize Model 4.A and the Firth's logistic regression approach utilizes Model 4.B.

$$\text{Model 4.A: } \log(\mathrm{E}[Y]) = \beta_0 + \beta_1 D,$$

$$\text{Model 4.B: } \operatorname{logit}(\mathrm{E}[D]) = \beta_0^* + \beta_1^* Y.$$

Each simulation scenario is replicated 1000 times.

All simulated data are analyzed with DESeq2, edgeR and Firth's logistic regression. The p-values from DESeq2, edgeR, and Firth's logistic regression are corrected for multiple tests using Bonferroni, Romano and Shaikh (RS), and Lehman and Romano (LR) procedures controlling FWER and Benjamini and Hochberg (BH), Benjamini and Yekutieli (BY), Storey and Tibshirani (ST), Benjamini, Krieger, and Yekutieli (BKY), and Blanchard and Roquain (BR) procedures controlling FDR at significance level of 0.05. Using the multiple testing corrected p-values from each multiple testing method, we identify Type-I errors and Type-II errors using the known DE status of genes in the simulation process. For multiple testing methods controlling the FWER, we calculate false positive rate (1 – specificity) and power (sensitivity; 1- false negative rate) at a significance level of 0.05. For multiple testing methods controlling the FDR, we compute FDR and power at a significance level of 0.05.

4.3 Results

4.3.1 False positive rates from multiple testing correction procedures controlling the FWER from simulated data

The false positive rates computed from simulated data sets with the proportion of null hypotheses equal to 100%, 95% and 75% are presented in Figure 4.1, Figure 4.2 and Figure 4.3, respectively. The false positive rates from edgeR are higher than from DESeq2 and Firth's logistic regression within the same multiple testing methods. The LR procedure has higher false positive rates than the other multiple testing correction methods within the same analysis methods. As sample size increases, the differences of false positive rates among analysis methods and among multiple testing methods decreases. In particular, with the sample size of five cases and five controls (Figure 4.1(A)), Firth's logistic regression did not identify any significant genes among null hypotheses in any of the simulated data sets. When sample size is small, the Firth's logistic regression is very conservative. Thus, Bonferroni, RS and LR procedures produced false positive rates equal to zero. Within the same multiple testing correction approach, the false positive rates from simulated data sets with different correlation structures are similar.

When the proportion of null hypotheses is decreased from 100% (Figure 4.1) to 75% (Figure 4.3), the false positive rates generally decrease. These decreases in the false positive rates are larger when sample size is small.

The false positive rates for different strengths of correlation with the sample size of five cases and five controls allowing no-correlation structures in DE genes analyzed with edgeR are presented in Table 4.2. The strength of correlation within the same correlation structure does not change false positive rates for edgeR, DESeq2, or Firth’s logistic regression, as shown in Table 4.2.

Comparison of false positive rates for the sample size of five cases and five controls analyzed with edgeR either with or without correlation structures in DE genes is shown in Table 4.3. The false positive rates for data sets with correlation structures in DE genes is similar to the false positive rates for data sets with no-correlation structures in DE genes. This similarity is also observed in the DESeq2 and Firth’s logistic regression results.

Figure 4.1 False positive rates with 100% null hypotheses

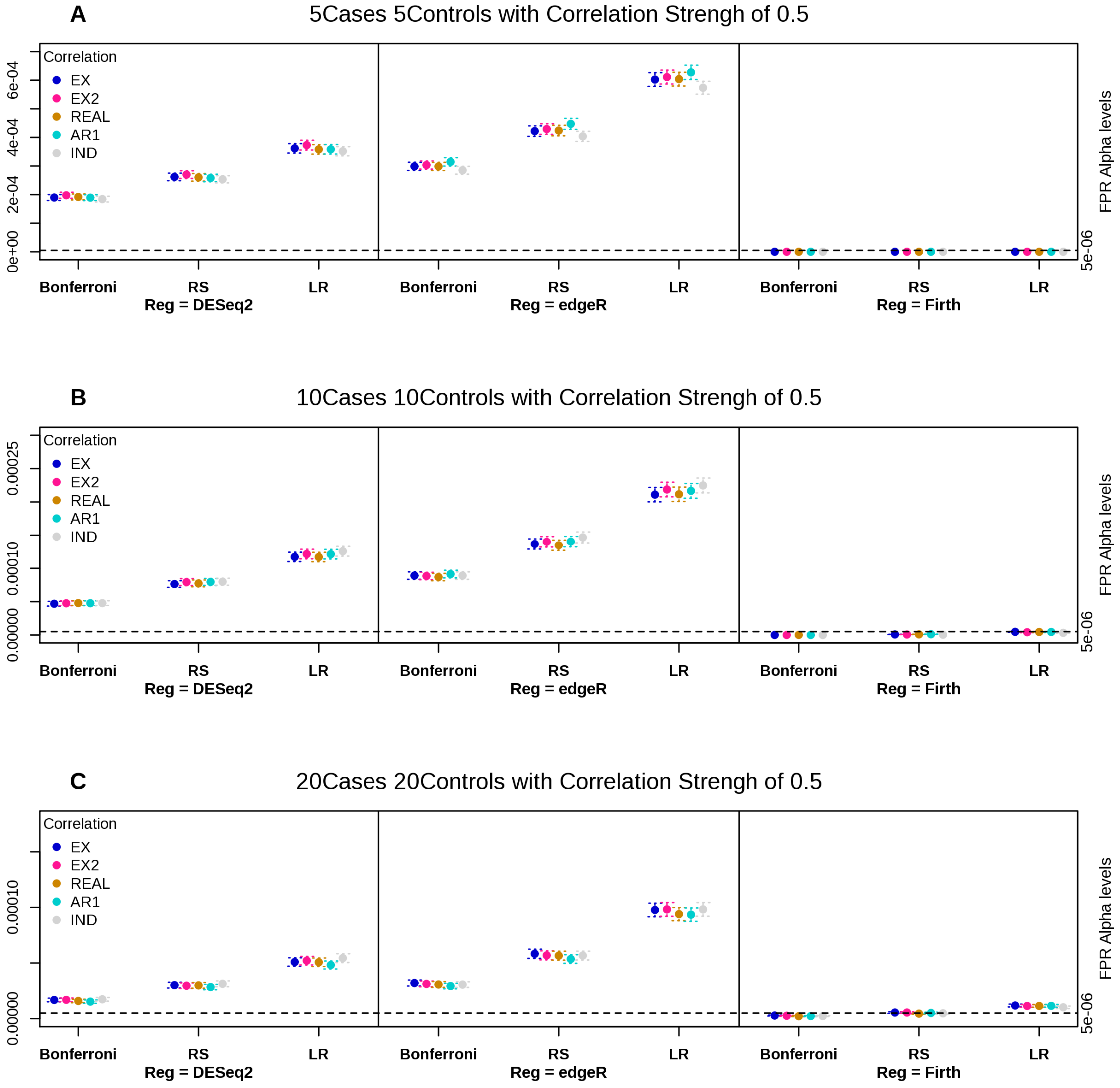
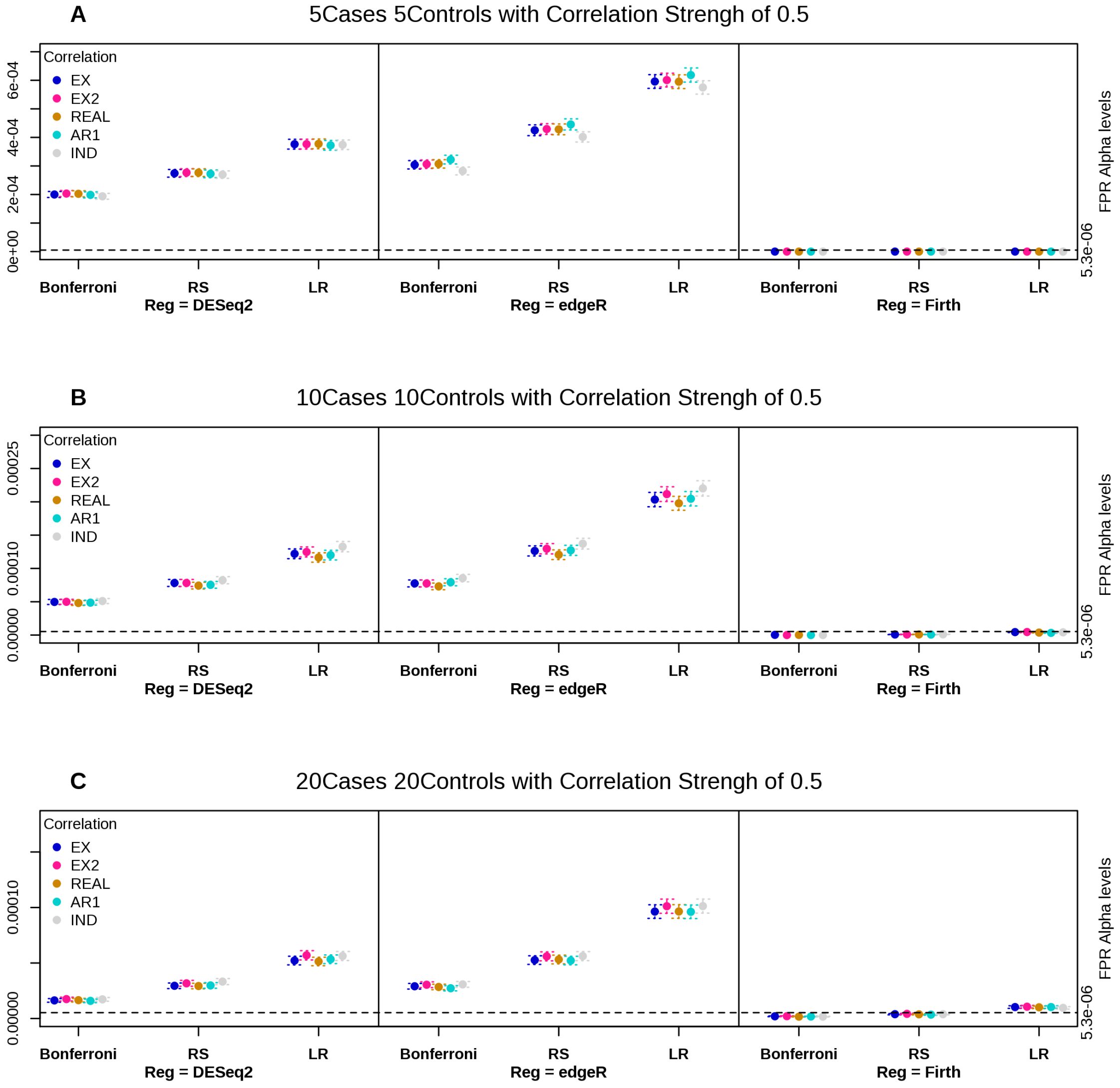


Figure 4.1 presents false positive rates from FWER methods. The analysis methods (DESeq2, edgeR, and Firth) are separated by vertical lines. Three FWER methods (Bonferroni, RS (Romano and Shaikh), LR (Lehman and Romano)) are placed within each analysis method. Colored dots represent exchangeable (EX), exchangeable2 (EX2), real, autoregressive1(AR1), and independent(IND) correlation structures in the null hypothesis. The dotted lines within each colored dot are the 95% confidence intervals of false positive rates. (A) presents the false positive rates for the sample size of five cases and five controls with correlation strength of 0.5, (B) presents the false positive rates for the sample size of 10 cases and 10 controls with correlation strength of 0.5, (C) presents the false positive rates for the sample size of 20 cases and 20 controls with correlation strength of 0.5

Figure 4.2 False positive rates with 95% null hypotheses


5Cases 5Controls with Correlation Strengh of 0.5

10Cases 10Controls with Correlation Strengh of 0.5

20Cases 20Controls with Correlation Strengh of 0.5

Figure 4.2 presents false positive rates from FWER methods. The analysis methods (DESeq2, edgeR, and Firth) are separated by vertical lines. Three FWER methods (Bonferroni, RS (Romano and Shaikh), LR (Lehman and Romano)) are placed within each analysis method. Colored dots represent exchangeable (EX), exchangeable2 (EX2), real, autoregressive1(AR1), and independent(IND) correlation structures in the null hypothesis. The dotted lines within each colored dot are the 95% confidence intervals of false positive rates. (A) presents the false positive rates for the sample size of five cases and five controls with correlation strength of 0.5, (B) presents the false positive rates for the sample size of 10 cases and 10 controls with correlation strength of 0.5, (C) presents the false positive rates for the sample size of 20 cases and 20 controls with correlation strength of 0.5

Figure 4.3 False positive rates with 75% null hypotheses

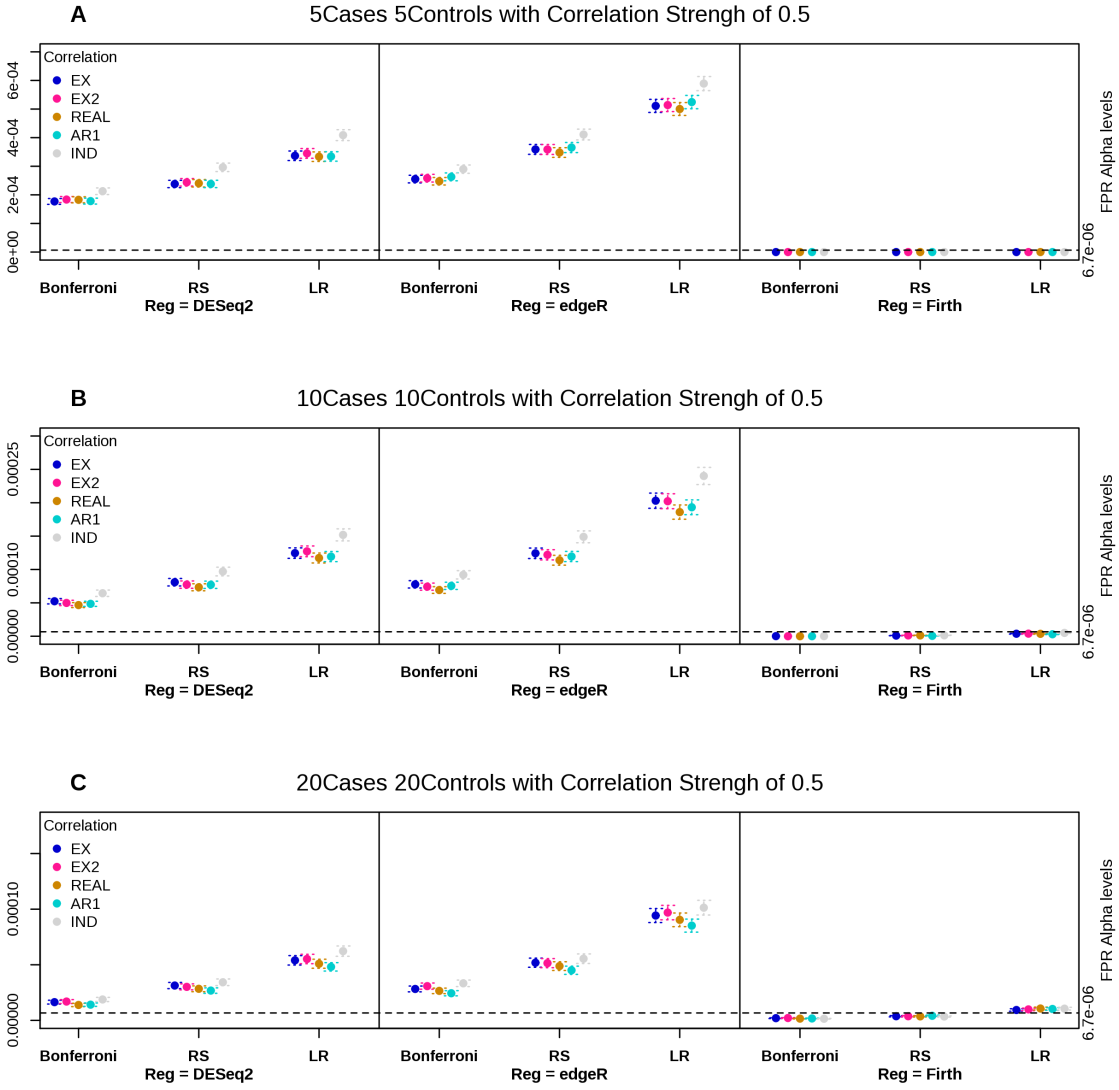


Figure 4.3 presents false positive rates of the FWER methods. The analysis methods (DESeq2, edgeR, and Firth) are separated by vertical lines. Three FWER methods (Bonferroni, RS (Romano and Shaikh), LR (Lehman and Romano)) are placed within each analysis method. Colored dots represent exchangeable (EX), exchangeable2 (EX2), real, autoregressive1(AR1), and independent(IND) correlation structures in the null hypothesis. The dotted lines within each colored dot are the 95% confidence intervals of false positive rates. (A) presents the false positive rates for a sample size of five cases and five controls with correlation strength of 0.5, (B) presents the false positive rates for the sample size of 10 cases and 10 controls with correlation strength of 0.5, (C) presents the false positive rates for the sample size of 20 cases and 20 controls with correlation strength of 0.5

Table 4.2 False positive rates for five cases and five controls from simulated data with no-correlation among differentially expressed genes based on analysis with edgeR.

Null	Type	Cor	Bonferroni	RS	LR
1	EX	0.3	0.00031	0.00044	0.00062
1	EX	0.5	0.0003	0.00042	0.0006
1	EX	0.75	0.00029	0.00041	0.00058
1	EX2	0.3	0.00031	0.00044	0.00062
1	EX2	0.5	0.0003	0.00043	0.00061
1	EX2	0.75	0.00029	0.00041	0.00058
1	AR1	0.3	0.00032	0.00045	0.00063
1	AR1	0.5	0.00031	0.00045	0.00063
1	AR1	0.75	0.00031	0.00043	0.00061
0.95	EX	0.3	0.00031	0.00044	0.0006
0.95	EX	0.5	0.0003	0.00043	0.0006
0.95	EX	0.75	0.0003	0.00042	0.00059
0.95	EX2	0.3	0.00032	0.00044	0.00061
0.95	EX2	0.5	0.00031	0.00043	0.0006
0.95	EX2	0.75	0.00029	0.00042	0.00059
0.95	AR1	0.3	0.00033	0.00045	0.00062
0.95	AR1	0.5	0.00032	0.00045	0.00062
0.95	AR1	0.75	0.00031	0.00044	0.00061
0.75	EX	0.3	0.00026	0.00036	0.00051
0.75	EX	0.5	0.00026	0.00036	0.00051
0.75	EX	0.75	0.00025	0.00035	0.00051
0.75	EX2	0.3	0.00026	0.00036	0.00052
0.75	EX2	0.5	0.00026	0.00036	0.00051
0.75	EX2	0.75	0.00025	0.00036	0.00052
0.75	AR1	0.3	0.00026	0.00037	0.00052
0.75	AR1	0.5	0.00026	0.00037	0.00052
0.75	AR1	0.75	0.00026	0.00036	0.00052

Null: Proportion of null hypothesis in a gene set, Type: Correlation structure types; exchangeable (EX), ecxchangeable2 (EX2), and autoregressive1(AR1), Cor: The strength of correlation in correlation structures. RS: Romano and Shaikh procedure, LR: Lehman and Romano procedure.

Table 4.3 False positive rates for five cases and five controls from simulated data based on analysis with edgeR.

DE-Cor	Null	Type	Cor	Bonferroni	RS	LR
No	0.95	EX	0.5	0.0003	0.00043	0.0006
	0.95	EX2	0.5	0.00031	0.00043	0.0006
	0.95	REAL	NA	0.00031	0.00043	0.0006
	0.95	AR1	0.5	0.00032	0.00045	0.00062
	0.95	IND	NA	0.00028	0.0004	0.00058
	0.75	EX	0.5	0.00026	0.00036	0.00051
	0.75	EX2	0.5	0.00026	0.00036	0.00051
	0.75	REAL	NA	0.00025	0.00035	0.0005
	0.75	AR1	0.5	0.00026	0.00037	0.00052
	0.75	IND	NA	0.00029	0.00041	0.00059
Yes	0.95	EX	0.5	0.00031	0.00043	0.0006
	0.95	EX2	0.5	0.00031	0.00043	0.0006
	0.95	REAL	NA	0.0003	0.00042	0.00059
	0.95	AR1	0.5	0.00032	0.00044	0.00062
	0.95	IND	NA	0.00029	0.0004	0.00057
	0.75	EX	0.5	0.00025	0.00034	0.00048
	0.75	EX2	0.5	0.00025	0.00035	0.0005
	0.75	REAL	NA	0.00024	0.00034	0.00048
	0.75	AR1	0.5	0.00025	0.00035	0.0005
	0.75	IND	NA	0.00028	0.00041	0.00058

DE-Cor: Presence of correlation in differentially expressed genes, Null: Proportion of null hypothesis in a gene set, Type: Correlation structure types; exchangeable (EX), ecxchangeable2 (EX2), real from Pickrell (REAL), autoregressive1(AR1) and independent (IND), Cor: The strength of correlation in correlation structures. RS: Romano and Shaikh procedure, LR: Lehman and Romano procedure.

4.3.2 FDRs from multiple testing correction procedures controlling FDR using simulated data.

The FDRs based on simulated data with 95% or 75% of the observations under the null hypotheses are presented in Figure 4.4 and Figure 4.5. The FDRs from analyses performed with edgeR are higher than FDRs from DESeq2 and Firth's logistic regression. In general, the BH, ST and BKY procedures produce similar FDRs, and the BY and BR procedures produce similar FDRs. However, the BH, ST, and BKY procedures have higher FDRs than the BY and BR procedures. FDRs based on data sets with different correlation structures within a multiple testing correction method are not distinctive enough to suggest that choice of multiple correction method should be influenced by correlation structure.

When the proportion of null hypotheses is 95% and the sample size for cases and controls is five (Figure 4.4(A)), the FDRs for the BY and BR procedures when using DESeq2 or edge R are close to the significance level of 0.05. In contrast, the BH, ST and BKY procedures have inflated false discovery rates when the sample size is small. Firth's logistic regression identifies no significant results with five cases and five controls.

When sample size increases (Figure 4.4(B) and (C)), the observed FDRs of the BH, ST, and BKY procedures are closer to of the specified 0.05 level. With increasing sample size, the BY and BR procedures become very conservative.

The FDRs of the BH, ST and BKY procedures computed from the DESeq2 method are closer to the specified FDR level of 0.05 than are the FDRs computed from the edgeR method. The FDRs from Firth’s logistic regression are conservative even with 20 cases and 20 controls.

FDRs for a sample size of five cases and five controls when there is no correlation among DE genes from analysis with edgeR are shown in Table 4.4. The strength of correlation within a correlation structure type does not notably affect false discovery rates for any of the analysis methods under our simulations. The assessment of FDRs for five cases and five controls from analysis with edgeR between correlation and no correlation among DE genes is presented in Table 4.5. The FDRs for correlation and no correlation among DE genes are very similar. This similarity is also found in DESeq2 and Firth’s logistic regression results.

Figure 4.4 False discovery rates with 95% null hypotheses

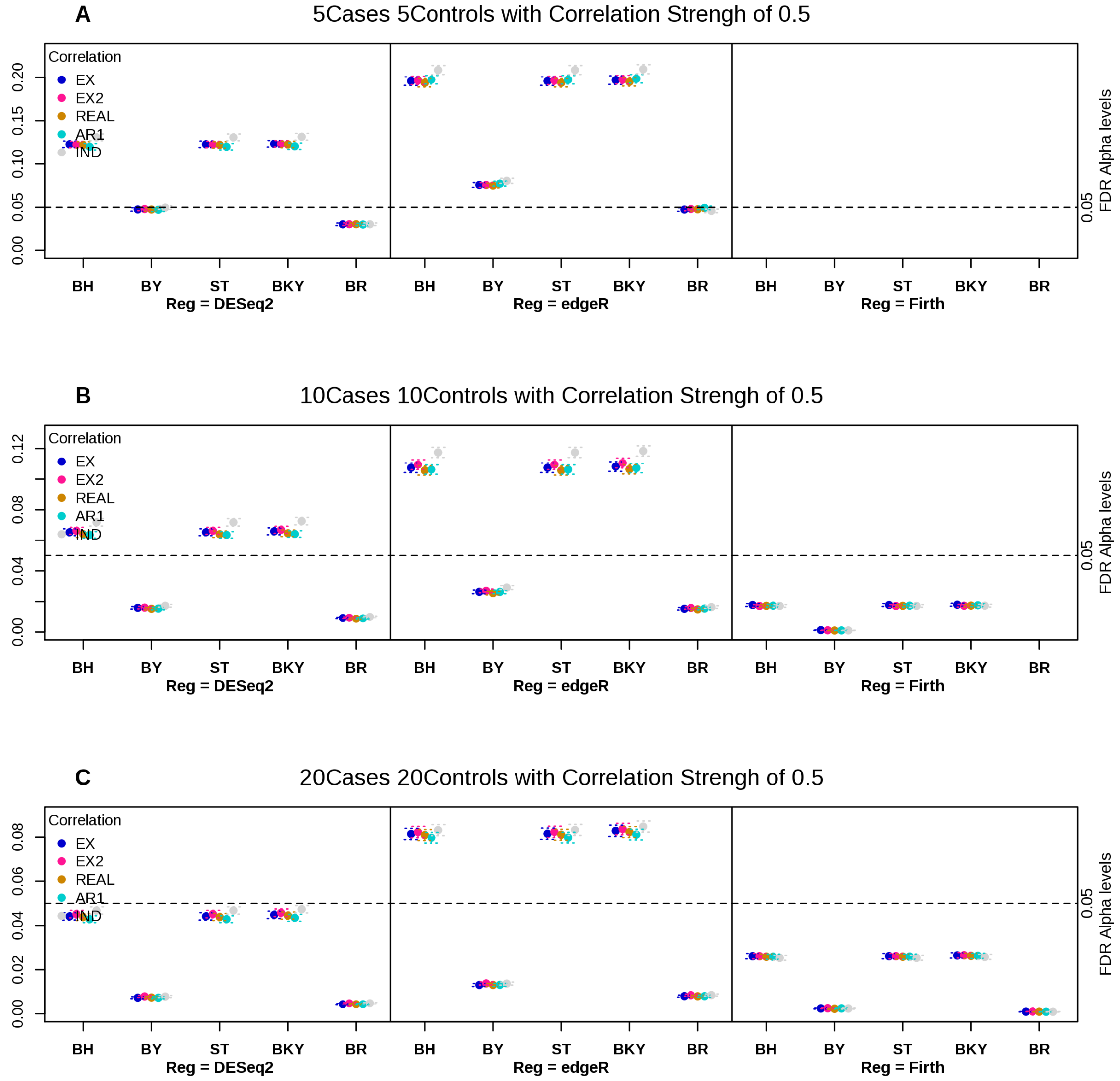


Figure 4.4 presents FDRs of the multiple testing correction methods controlling the FDR. The analysis methods (DESeq2, edgeR, and Firth) are separated by vertical lines. Five FDR methods (BH (Benjamini and Hochberg), BY (Benjamini and Yekutieli), ST (Story and Tibshirani), BKY (Benjamini, Krieger, and Yekutieli), BR (Blanchard and Roquain)) are placed within each analysis method. Colored dots represent exchangeable (EX), exchangeable2 (EX2), real, autoregressive1(AR1), and independent(IND) correlation structures in the null hypothesis. The dotted lines within each colored dot are the 95% confidence intervals of false discovery rates. (A) presents the false discovery rates for the sample size of five cases and five controls with correlation strength of 0.5, (B) presents the false discovery rates for the sample size of 10 cases and 10 controls with correlation strength of 0.5, (C) presents the false discovery rates for the sample size of 20 cases and 20 controls with correlation strength of 0.5

Figure 4.5 False discovery rates with 75% null hypotheses

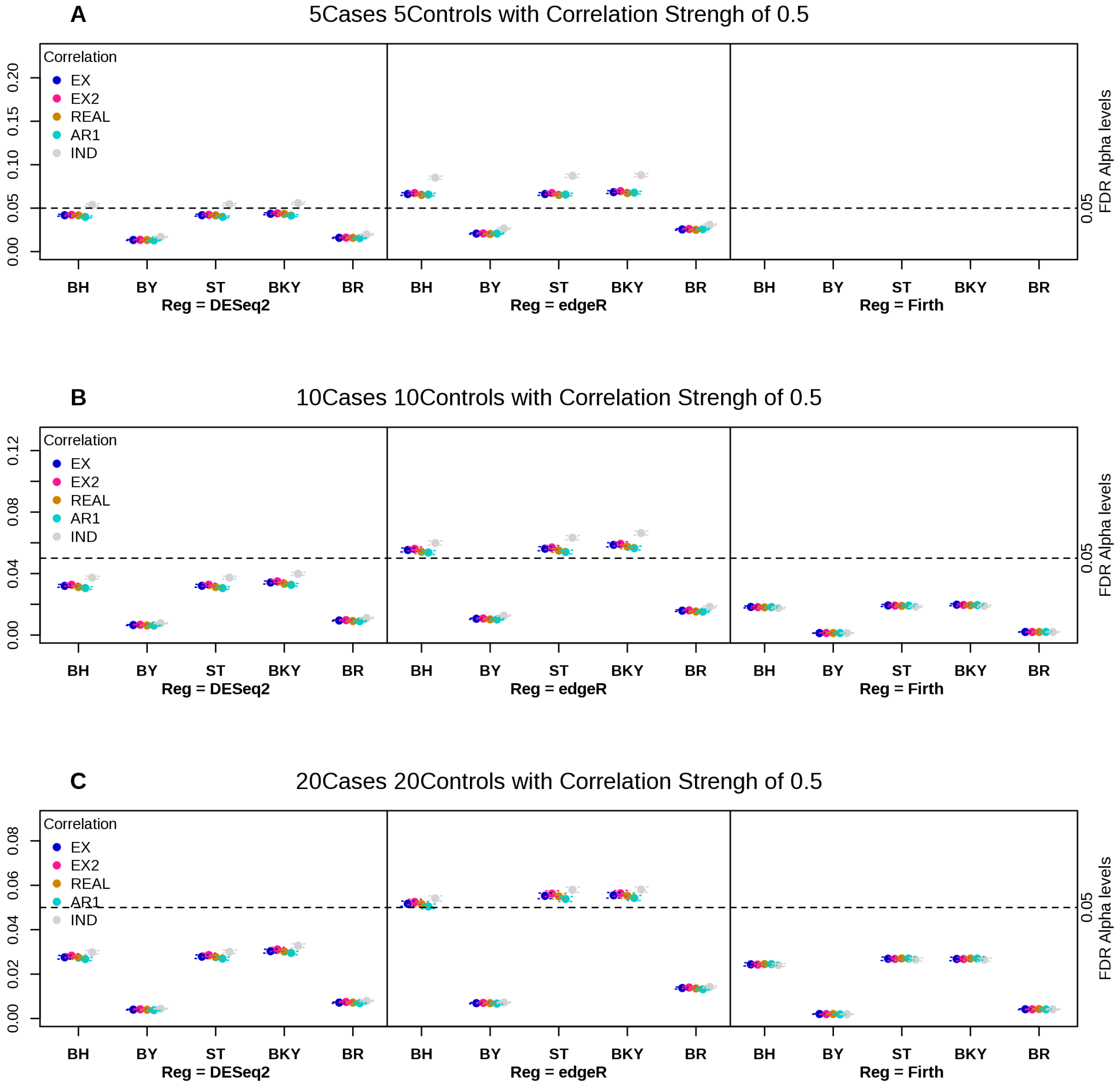


Figure 4.5 presents FDRs of the multiple testing correction methods controlling the FDR. The analysis methods (DESeq2, edgeR, and Firth) are separated by vertical lines. Five FDR methods (BH (Benjamini and Hochberg), BY (Benjamini and Yekutieli), ST (Story and Tibshirani), BKY (Benjamini, Krieger, and Yekutieli), BR (Blanchard and Roquain)) are placed within each analysis method. Colored dots represent exchangeable (EX), exchangeable2 (EX2), real, autoregressive1(AR1), and independent(IND) correlation structures in the null hypothesis. The dotted lines within each colored dot are the 75% confidence intervals of false discovery rates. (A) presents the false discovery rates for the sample size of five cases and five controls with correlation strength of 0.5, (B) presents the false discovery rates for the sample size of 10 cases and 10 controls with correlation strength of 0.5, (C) presents the false discovery rates for the sample size of 20 cases and 20 controls with correlation strength of 0.5

Table 4.4 False discovery rates for five cases and five controls from simulated data with no-correlation among differentially expressed genes based on analysis with edgeR

Null	Type	Cor	BH	BY	ST	BKY	BR
0.95	EX	0.3	0.196	0.076	0.196	0.197	0.048
0.95	EX	0.5	0.196	0.076	0.196	0.197	0.047
0.95	EX	0.75	0.196	0.075	0.196	0.197	0.047
0.95	EX2	0.3	0.197	0.077	0.197	0.198	0.048
0.95	EX2	0.5	0.197	0.076	0.197	0.198	0.048
0.95	EX2	0.75	0.199	0.075	0.199	0.200	0.047
0.95	AR1	0.3	0.198	0.078	0.198	0.199	0.050
0.95	AR1	0.5	0.197	0.077	0.197	0.198	0.049
0.95	AR1	0.75	0.197	0.077	0.197	0.198	0.048
0.75	EX	0.3	0.065	0.021	0.065	0.068	0.026
0.75	EX	0.5	0.066	0.021	0.066	0.068	0.025
0.75	EX	0.75	0.069	0.021	0.069	0.071	0.026
0.75	EX2	0.3	0.066	0.021	0.066	0.068	0.026
0.75	EX2	0.5	0.067	0.021	0.067	0.070	0.026
0.75	EX2	0.75	0.072	0.022	0.072	0.074	0.027
0.75	AR1	0.3	0.066	0.021	0.066	0.068	0.026
0.75	AR1	0.5	0.066	0.021	0.066	0.068	0.026
0.75	AR1	0.75	0.066	0.021	0.066	0.068	0.026

Null: Proportion of null hypothesis in a gene set, Type: Correlation structure types; exchangeable (EX), ecxchangeable2 (EX2), and autoregressive1(AR1), Cor: The strength of correlation in correlation structures. BH: Benjamini and Hochberg, BY: Benjamini and Yekutieli, ST: Story and Tibshirani, BKY: Benjamini, Krieger, and Yekutieli, BR: Blanchard and Roquain.

Table 4.5 False discovery rates for five cases and five controls from simulated data based on analysis with edgeR

DE-Cor	Null	Type	Cor	BH	BY	ST	BKY	BR
No	0.95	EX	0.5	0.196	0.076	0.196	0.197	0.047
	0.95	EX2	0.5	0.197	0.076	0.197	0.198	0.048
	0.95	REAL	NA	0.194	0.075	0.194	0.195	0.048
	0.95	AR1	0.5	0.197	0.077	0.197	0.198	0.049
	0.95	IND	NA	0.209	0.080	0.209	0.210	0.046
	0.75	EX	0.5	0.066	0.021	0.066	0.068	0.025
	0.75	EX2	0.5	0.067	0.021	0.067	0.070	0.026
	0.75	REAL	NA	0.065	0.020	0.065	0.067	0.025
	0.75	AR1	0.5	0.066	0.021	0.066	0.068	0.026
	0.75	IND	NA	0.085	0.027	0.087	0.088	0.031
Yes	0.95	EX	0.5	0.196	0.076	0.196	0.197	0.048
	0.95	EX2	0.5	0.197	0.076	0.197	0.198	0.048
	0.95	REAL	NA	0.194	0.075	0.194	0.196	0.048
	0.95	AR1	0.5	0.199	0.077	0.199	0.200	0.050
	0.95	IND	NA	0.209	0.081	0.209	0.210	0.047
	0.75	EX	0.5	0.066	0.02	0.066	0.068	0.025
	0.75	EX2	0.5	0.066	0.021	0.066	0.069	0.025
	0.75	REAL	NA	0.065	0.020	0.065	0.067	0.025
	0.75	AR1	0.5	0.065	0.020	0.065	0.067	0.025
	0.75	IND	NA	0.085	0.026	0.087	0.088	0.030

DE-Cor: Presence of correlation in differentially expressed genes, Null: Proportion of null hypothesis in a gene set, Type: Correlation structure types; exchangeable (EX), ecxchangeable2 (EX2), real from Pickrell (REAL), autoregressive1(AR1) and independent (IND), Cor: The strength of correlation in correlation structures. BH: Benjamini and Hochberg, BY: Benjamini and Yekutieli, ST: Story and Tibshirani, BKY: Benjamini, Krieger, and Yekutieli, BR: Blanchard and Roquain.

4.3.3 Power comparison among the multiple testing correction methods controlling FWER based on simulated data.

Power for different multiple testing methods controlling FWER applied to simulated data sets with 75% null hypotheses is shown in Figure 4.6. The power from DEseq2 and from edgeR are comparable, and are higher than the power from Firth’s logistic regression. Although power using the LR procedure is greater than RS and Bonferroni, with Bonferroni procedure showing the lowest power, the differences in power among these three methods are very small except for Firth’s logistic regression with 10 cases and 10 controls (Figure 4.6(B)). As sample size increases, power increases. Power for 95% null hypotheses is similar to power with 75% null hypotheses (Table 4.6). Strength of correlation (Table 4.6) and presence of correlation structures in DE genes (Table 4.7) do not influence power.

Figure 4.6 Power with proportion of null hypotheses equal to 75% using the multiple testing correction methods controlling FWER

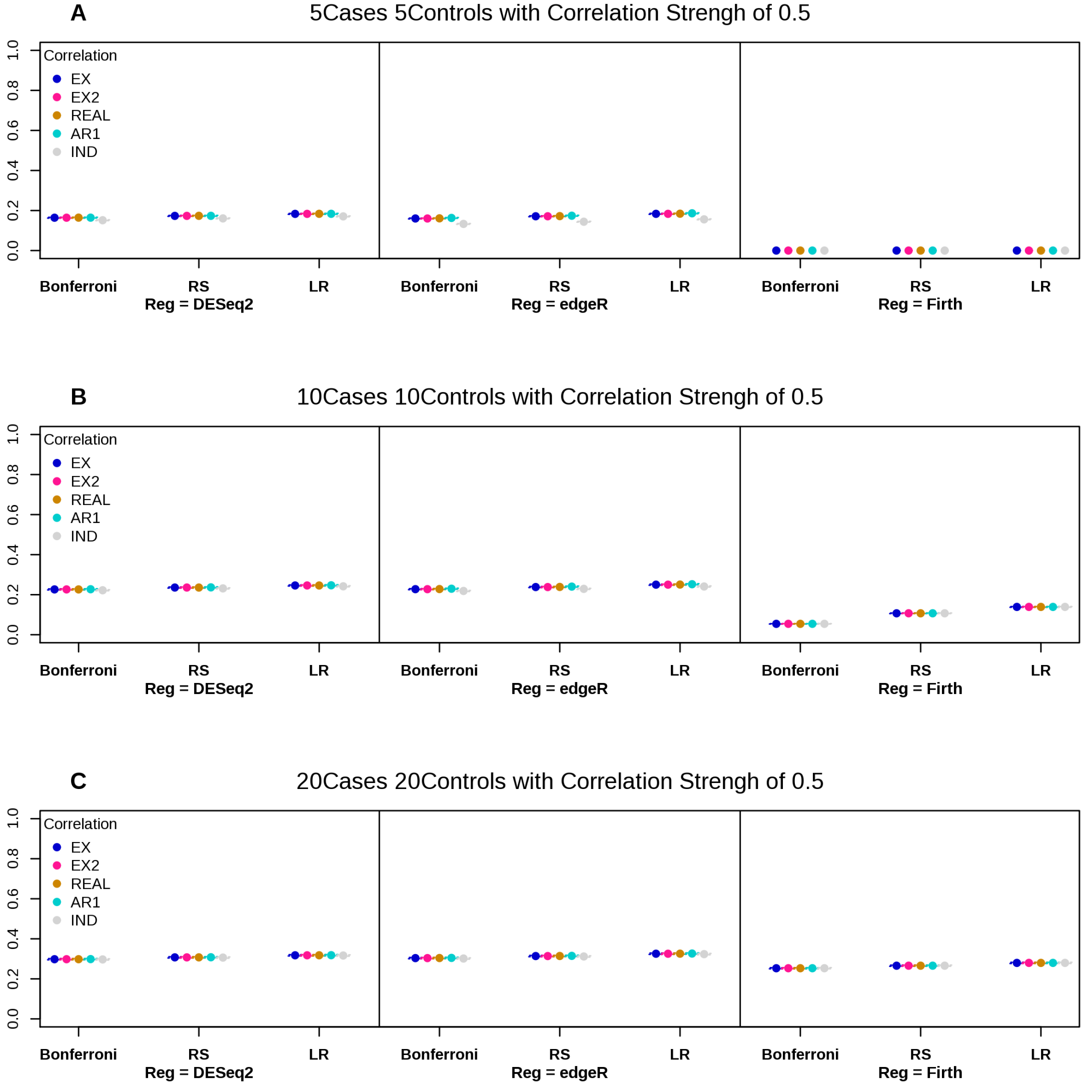


Figure 4.6 presents power from the multiple testing correction methods controlling FWER. The analysis methods (DESeq2, edgeR, and Firth) are separated by vertical lines. Three FWER methods (Bonferroni, RS (Romano and Shaikh), LR (Lehman and Romano)) are placed within each analysis method. Colored dots represent exchangeable (EX), exchangeable2 (EX2), real, autoregressive1(AR1), and independent(IND) correlation structures in the null hypothesis. The dotted lines within each colored dot are the 95% confidence intervals of power(A) presents the power for the sample size of five cases and five controls with correlation strength of 0.5, (B) presents the power for the sample size of 10 cases and 10 controls with correlation strength of 0.5, (C) presents the power for the sample size of 20 cases and 20 controls with correlation strength of 0.5

Table 4.6 Power for sample size of five cases and five controls with no-correlation among the differentially expressed genes based on analysis using edgeR and the multiple testing correction methods controlling FWER

Null	Type	Cor	Bonferroni	RS	LR
0.95	EX	0.3	0.163	0.174	0.185
0.95	EX	0.5	0.161	0.172	0.183
0.95	EX	0.75	0.156	0.167	0.178
0.95	EX2	0.3	0.163	0.174	0.186
0.95	EX2	0.5	0.161	0.172	0.183
0.95	EX2	0.75	0.155	0.166	0.177
0.95	AR1	0.3	0.164	0.175	0.187
0.95	AR1	0.5	0.164	0.175	0.186
0.95	AR1	0.75	0.162	0.173	0.185
0.75	EX	0.3	0.162	0.173	0.185
0.75	EX	0.5	0.160	0.171	0.183
0.75	EX	0.75	0.155	0.166	0.178
0.75	EX2	0.3	0.162	0.173	0.186
0.75	EX2	0.5	0.160	0.171	0.183
0.75	EX2	0.75	0.154	0.165	0.177
0.75	AR1	0.3	0.163	0.174	0.187
0.75	AR1	0.5	0.163	0.174	0.186
0.75	AR1	0.75	0.161	0.172	0.185

Null: Proportion of null hypothesis in a gene set, Type: Correlation structure types; exchangeable (EX), ecxchangeable2 (EX2), and autoregressive1(AR1), Cor: The strength of correlation in correlation structures. RS: Romano and Shaikh procedure, LR: Lehman and Romano procedure.

Table 4.7 Power for the sample size of five cases and five controls based on analysis using edgeR results and the multiple testing correction methods controlling FWER

DE-Cor	Null	Type	Cor	Bonferroni	RS	LR
No	0.95	EX	0.5	0.161	0.172	0.183
	0.95	EX2	0.5	0.161	0.172	0.183
	0.95	REAL	NA	0.162	0.172	0.184
	0.95	AR1	0.5	0.164	0.175	0.186
	0.95	IND	NA	0.134	0.145	0.156
	0.75	EX	0.5	0.160	0.171	0.183
	0.75	EX2	0.5	0.160	0.171	0.183
	0.75	REAL	NA	0.161	0.172	0.184
	0.75	AR1	0.5	0.163	0.174	0.186
	0.75	IND	NA	0.133	0.144	0.156
Yes	0.95	EX	0.5	0.159	0.170	0.181
	0.95	EX2	0.5	0.159	0.170	0.181
	0.95	REAL	NA	0.159	0.170	0.182
	0.95	AR1	0.5	0.161	0.172	0.184
	0.95	IND	NA	0.132	0.142	0.154
	0.75	EX	0.5	0.159	0.170	0.182
	0.75	EX2	0.5	0.159	0.170	0.182
	0.75	REAL	NA	0.160	0.171	0.183
	0.75	AR1	0.5	0.162	0.173	0.185
	0.75	IND	NA	0.132	0.143	0.155

DE-Cor: Presence of correlation in differentially expressed genes, Null: Proportion of null hypothesis in a gene set, Type: Correlation structure types; exchangeable (EX), ecxchangeable2 (EX2), real from Pickrell (REAL), autoregressive1(AR1) and independent (IND), Cor: The strength of correlation in correlation structures. BH: Benjamini and Hochberg, BY: Benjamini and Yekutieli, ST: Story and Tibshirani, BKY: Benjamini, Krieger, and Yekutieli, BR: Blanchard and Roquain.

4.3.4 Power comparison among the multiple testing correction methods controlling FDR based on simulated data.

Power for the multiple testing methods controlling FDR based on for simulated data sets with 95% and 75% null hypotheses are shown in Figure 4.7 and Figure 4.8. The power from analysis with DEseq2 and from edgeR are comparable. Although they are higher than the power from Firth's logistic regression, as sample size increases, this differences in power among analysis methods decreases. The BH, ST, BKY procedures have similar power and are more powerful than BY and BR. As sample size increases, the differences in power among the multiple testing methods decreases. In general, large sample size increases power across all analysis and multiple testing methods. When the proportion of null hypotheses in a data set decreases from 95%(Figure 4.7) to 75%(Figure 4.8), power is increased. Strength of correlation (Table 4.8) and presence of correlation structures in DE genes (Table 4.9) do not influence the power.

Figure 4.7 Power with proportion of null hypotheses equal to 95% using FDR methods

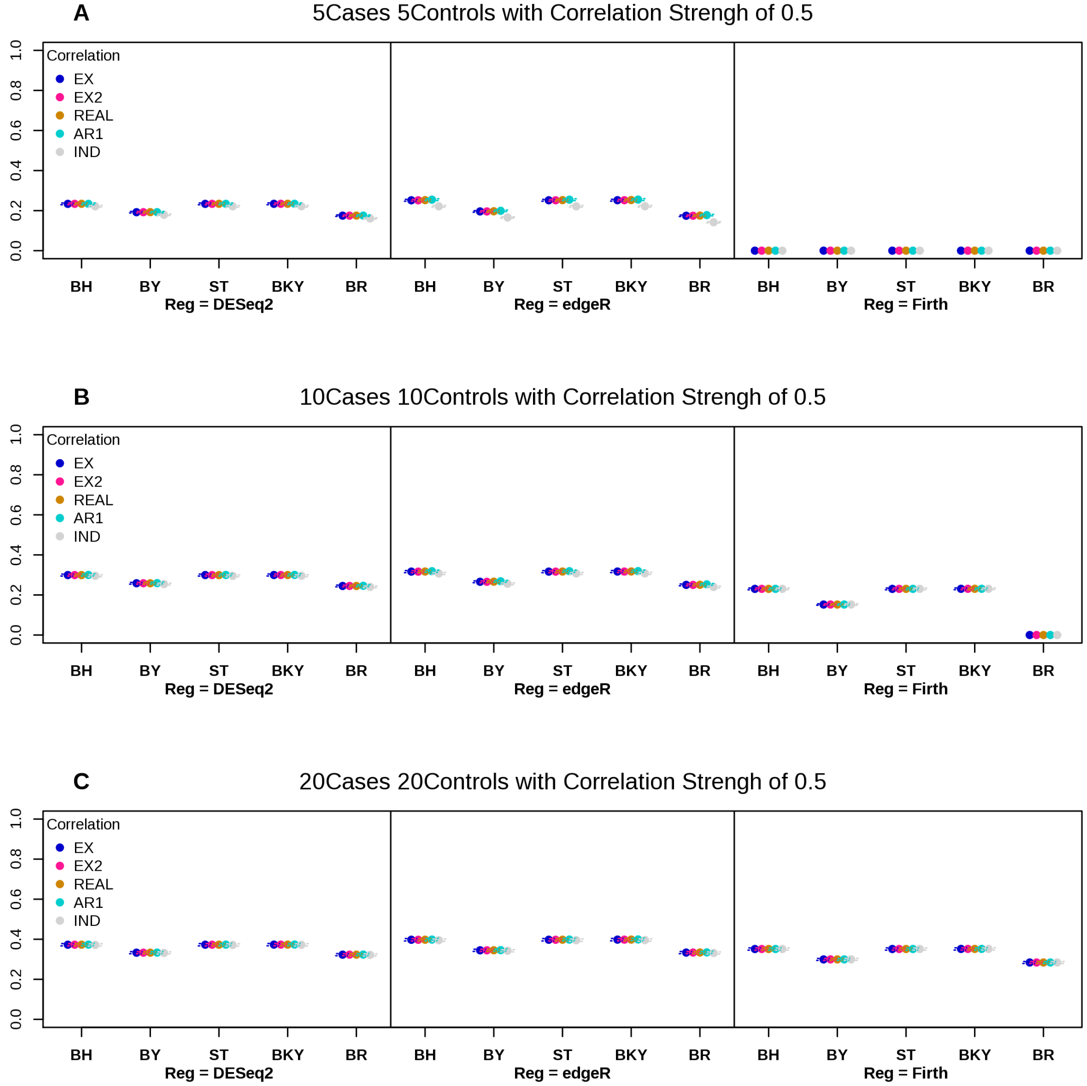


Figure 4.7 presents power from the multiple testing correction methods controlling FDR. The analysis methods (DESeq2, edgeR, and Firth) are separated by vertical lines. Five FDR methods (BH (Benjamini and Hochberg), BY (Benjamini and Yekutieli), ST (Story and Tibshirani), BKY (Benjamini, Krieger, and Yekutieli), BR (Blanchard and Roquain)) are placed within each analysis method. Colored dots represent exchangeable (EX), exchangeable2 (EX2), real, autoregressive1(AR1), and independent(IND) correlation structures in the null hypothesis. The dotted lines within each colored dot are the 95% confidence intervals of power. (A) presents power from the sample size of five cases and five controls with correlation strength of 0.5, (B) presents power from the sample size of 10 cases and 10 controls with correlation strength of 0.5, (C) presents power from the sample size of 20 cases and 20 controls with correlation strength of 0.5

Figure 4.8 Power with proportion of null hypotheses equal to 75% using FDR methods

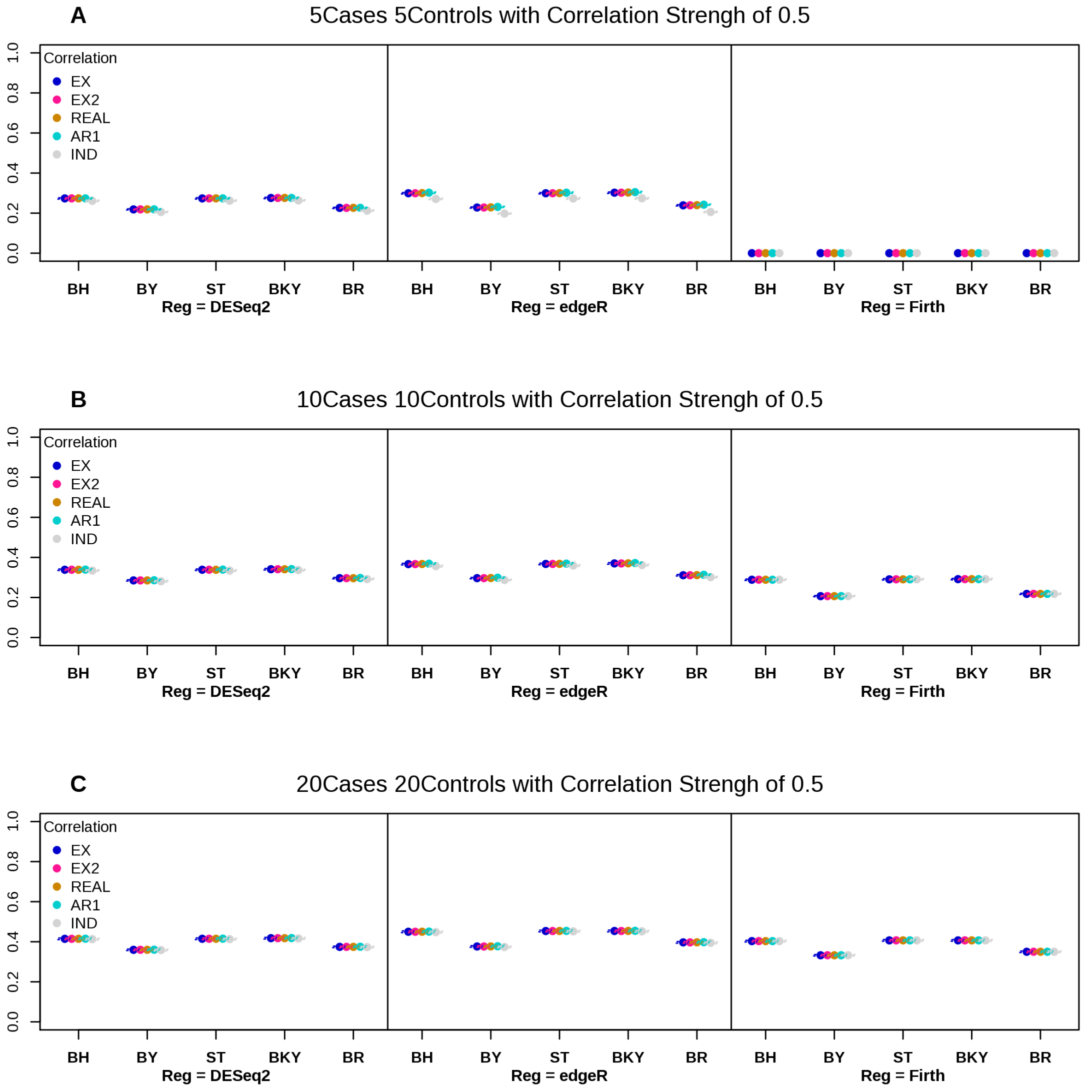


Figure 4.8 presents power from the multiple testing correction methods controlling FDR. The analysis methods (DESeq2, edgeR, and Firth) are separated by vertical lines. Five FDR methods (BH (Benjamini and Hochberg), BY (Benjamini and Yekutieli), ST (Story and Tibshirani), BKY (Benjamini, Krieger, and Yekutieli), BR (Blanchard and Roquain)) are placed within each analysis method. Colored dots represent exchangeable (EX), exchangeable2 (EX2), real, autoregressive1(AR1), and independent(IND) correlation structures in the null hypothesis. The dotted lines within each colored dot are the 95% confidence intervals of power. (A) presents power from the sample size of five cases and five controls with correlation strength of 0.5, (B) presents power from the sample size of 10 cases and 10 controls with correlation strength of 0.5, (C) presents power from the sample size of 20 cases and 20 controls with correlation strength of 0.5

Table 4.8 Power for the sample size of five cases and five controls in no-correlation in differentially expressed genes based on analysis using edgeR and the multiple testing correction methods controlling FDR

Null	Type	Cor	BH	BY	ST	BKY	BR
0.95	EX	0.3	0.254	0.198	0.254	0.254	0.177
0.95	EX	0.5	0.252	0.196	0.252	0.252	0.174
0.95	EX	0.75	0.246	0.190	0.246	0.246	0.168
0.95	EX2	0.3	0.254	0.198	0.254	0.254	0.177
0.95	EX2	0.5	0.252	0.196	0.252	0.252	0.174
0.95	EX2	0.75	0.245	0.189	0.245	0.245	0.167
0.95	AR1	0.3	0.255	0.200	0.255	0.256	0.178
0.95	AR1	0.5	0.255	0.199	0.255	0.255	0.178
0.95	AR1	0.75	0.253	0.197	0.253	0.253	0.176
0.75	EX	0.3	0.301	0.230	0.301	0.304	0.241
0.75	EX	0.5	0.299	0.228	0.299	0.302	0.239
0.75	EX	0.75	0.294	0.222	0.294	0.297	0.233
0.75	EX2	0.3	0.301	0.230	0.301	0.304	0.241
0.75	EX2	0.5	0.299	0.228	0.299	0.302	0.239
0.75	EX2	0.75	0.293	0.221	0.293	0.296	0.232
0.75	AR1	0.3	0.302	0.232	0.302	0.305	0.243
0.75	AR1	0.5	0.302	0.231	0.302	0.305	0.242
0.75	AR1	0.75	0.300	0.229	0.300	0.303	0.240

Null: Proportion of null hypothesis in a gene set, Type: Correlation structure types; exchangeable (EX), ecxchangeable2 (EX2), and autoregressive1(AR1), Cor: The strength of correlation in correlation structures. BH: Benjamini and Hochberg, BY: Benjamini and Yekutieli, ST: Story and Tibshirani, BKY: Benjamini, Krieger, and Yekutieli, BR: Blanchard and Roquain.

Table 4.9 Power for the sample size of five cases and five controls based on analysis using edgeR and the multiple testing correction methods controlling FDR

DE-Cor	Null	Type	Cor	BH	BY	ST	BKY	BR
No	0.95	EX	0.5	0.252	0.196	0.252	0.252	0.174
	0.95	EX2	0.5	0.252	0.196	0.252	0.252	0.174
	0.95	REALCOR	NA	0.252	0.197	0.252	0.253	0.175
	0.95	AR1	0.5	0.255	0.199	0.255	0.255	0.178
	0.95	IND	NA	0.221	0.166	0.221	0.222	0.142
	0.75	EX	0.5	0.299	0.228	0.299	0.302	0.239
	0.75	EX2	0.5	0.299	0.228	0.299	0.302	0.239
	0.75	REALCOR	NA	0.3	0.229	0.3	0.303	0.24
	0.75	AR1	0.5	0.302	0.231	0.302	0.305	0.242
	0.75	IND	NA	0.271	0.198	0.273	0.274	0.206
Yes	0.95	EX	0.5	0.249	0.193	0.249	0.249	0.172
	0.95	EX2	0.5	0.249	0.193	0.249	0.25	0.172
	0.95	REALCOR	NA	0.25	0.194	0.25	0.25	0.172
	0.95	AR1	0.5	0.252	0.196	0.252	0.253	0.175
	0.95	IND	NA	0.219	0.163	0.219	0.219	0.139
	0.75	EX	0.5	0.298	0.226	0.298	0.3	0.237
	0.75	EX2	0.5	0.298	0.226	0.298	0.3	0.237
	0.75	REALCOR	NA	0.298	0.227	0.298	0.301	0.238
	0.75	AR1	0.5	0.3	0.229	0.3	0.303	0.24
	0.75	IND	NA	0.269	0.196	0.271	0.272	0.204

DE-Cor: Presence of correlation in differentially expressed genes, Null: Proportion of null hypothesis in a gene set, Type: Correlation structure types; exchangeable (EX), ecxchangeable2 (EX2), real from Pickrell (REAL), autoregressive1(AR1) and independent (IND), Cor: The strength of correlation in correlation structures. BH: Benjamini and Hochberg, BY: Benjamini and Yekutieli, ST: Story and Tibshirani, BKY: Benjamini, Krieger, and Yekutieli, BR: Blanchard and Roquain.

4.4 Discussion

In this chapter, we compare multiple testing correction methods that control either FWER or FDR. Because the expression of many genes measured in RNA-Seq data are correlated, we simulate correlations among expression values of genes that are not differentially expressed. We demonstrate that power and type I error or false positive rates do not differ among various correlation and independence scenarios. Moreover, we find the strength of correlation among genes does not have an impact on performance.

Compared with multiple testing correction methods controlling FWER, the LR procedure, which controls FDR, has higher false positive rates than the Bonferroni and RS procedures. However, the differences in false positive rates among FWER methods are small. When the proportion of null hypotheses in a data set is decreased, false positive rates also decrease. Interestingly, the proportion of null hypotheses does not influence power. Although the LR procedure has slightly greater power than the Bonferroni and RS procedure, the differences in power are not substantive.

Multiple testing correction procedures imposing FDR, the BH, ST and BKY procedures produce similar FDRs and power. These three procedures have higher FDRs and power than the BY and BR procedures under our scenarios. The proportion of null hypotheses plays a critical role in the multiple testing correction methods controlling FDR. As the proportion of null hypotheses decreases, false discovery rates decrease, but power increases.

We also notice that although the RS and LR procedures were developed to control false positive rates and to increase power under any correlation structure, these procedures are not as powerful as other FDR methods. However, false positive rates and power do not vary much with the proportion of null hypotheses in the multiple testing correction methods controlling FWER. FDRs and power do vary with the proportion of null hypotheses in the multiple testing correction methods controlling FDR. The proportion of null hypotheses can differ widely depending on the disease or tissue. Knowing the proportion of null hypotheses in a RNA-Seq data set can be an important factor for interpreting results from the multiple testing correction methods controlling FDR.

It is important for researchers to be aware of the study designs, RNA-Seq data sets and the consequences of applying specific regression and multiple testing correction methods. For example, although the BY procedure is known to be a conservative method, the BY procedure controls the FDR well with five cases and five controls when the proportion of null hypothesis is equal to 95% using DESeq2 (Figure 4.4(A)).In contrast, when the proportion of null hypotheses is equal to 75% for the same sample size and analysis method, the BY procedure produces very conservative FDRs (Figure 4.5(A)). Researchers may seek a conservative false positive rate for their validation studies. Then, the LR procedure, which showed consistent power regardless of proportion of null hypotheses, is an attractive choice together with Firth’s logistic regression. Equivalently, the BH, ST and BKY procedures from Firth’s logistic regression provide conservative FDRs with reasonable power.

Chapter 5 Summary and future work

5.1 Summary

Exploration of statistical methodology has been critical for gene expression studies. This is particularly true for next generation sequencing technology in which the expression data generated are count data rather than quantitative measurements. Active statistical research in this field significantly improves the capability of detecting truly differentially expressed transcripts.

In this dissertation, we suggest an alternative statistical approach, and we evaluate the effect of covariates and the effect of correction structures in gene expression studies. In Chapter 2, we recommend the use of Firth's logistic regression to analyze RNA-Seq data in case-control studies. Because estimation of the dispersion parameter for each gene is not necessary in this approach, Firth's logistic regression provides a concise statistical inference process and reduces false positives from inaccurately estimated dispersion parameters in the negative binomial framework. Future work related to Firth's logistic regression in the RNA-Seq context will involve generating a genomic risk score that combines risks of multiple genes into a single variable. This genomic risk score may improve discrimination and calibration. In Chapter 3, we evaluate the effect of non-predictive covariates in negative binomial models and the effect of non-confounding predictive covariates in Firth's logistic models. We suggest that RNA-Seq data should be analyzed with a parsimonious model using Firth's logistic regression. When odds ratios between covariates and case-control status are moderate, Firth’s logistic regression is robust to the increase in number of non-confounding predictive covariates in a model. Because including a confounder in the model results in a more accurate model, we will explore a new algorithm that identifies relationships between covariates and genes, and then generates gene-specific models. Comparing performance of a conventional model and gene-specific models may underline the importance of precise modeling. In Chapter 4, we compare performance of multiple testing correction methods that control FWER or FDR. Although correlation structures under null hypotheses follow the negative binomial distribution and thus do not have a huge impact on performance of multiple testing correction methods, this study reveals that understanding study design, RNA-Seq data, and the expected consequence of analysis methods and multiple testing methods is imperative for RNA-Seq studies. This prior knowledge significantly contributes to the identification of an appropriate statistical method in gene expression studies.

In conclusion, we investigate analysis methods (Chapter 2), analysis models (Chapter 3), and multiple testing methods (Chapter 4) of RNA-Seq studies. We believe our conclusions and suggestions will enhance gene expression studies (Section 1.1) and influence related statistical areas including count data analysis, covariate analysis and correlated data analysis.

BIBLIOGRAPHY

Albert, A., and J A. Aanderson. 1984. “On the Existence of Maximum Likelihood Estimates in Logistic Regression Models.” *Biometrika* 71 (1) (April 1): 1–10. doi:10.1093/biomet/71.1.1.

Allison, Paul D. 2012. *Logistic Regression Using SAS: Theory and Application*. Second. SAS Institute.

Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology* 11 (10): R106. doi:10.1186/gb-2010-11-10-r106.

Arzberger, Thomas, Klaus Krampfl, Susanne Leimgruber, and Adolf Weindl. 1997. “Changes of NMDA Receptor Subunit (NR1, NR2B) and Glutamate Transporter (GLT1) mRNA Expression in Huntington’s Disease—An In Situ Hybridization Study.” *Journal of Neuropathology and Experimental Neurology* 56 (4) (April): 440–454. doi:10.1097/00005072-199704000-00013.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

Benjamini, Yoav, A. M. Krieger, and D. Yekutieli. 2006. “Adaptive Linear Step-up Procedures That Control the False Discovery Rate.” *Biometrika* 93 (3) (September 1): 491–507. doi:10.1093/biomet/93.3.491.

Berger, James, and L. Mark Berliner. 1986. “Robust Bayes and Empirical Bayes Analysis with $_\epsilon$-Contaminated Priors.” *The Annals of Statistics* 14 (2) (June 1): 461–486. doi:10.1214/aos/1176349933.

Black, M A. 2004. “A Note on the Adaptive Control of False Discovery Rates.” *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 66 (2) (January 1): 297–304. doi:10.2307/3647526.

Blanchard, Gilles, and Etienne Roquain. 2008. “Two Simple Sufficient Conditions for FDR Control.” *Electronic Journal of Statistics* 2: 963–992. doi:10.1214/08-EJS180.

Bonferroni, Carlo Emilio. 1936. “Teoria Statistica Delle Classi E Calcolo Delle Probabilita.” *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche E Commerciali Di Firenze* 8: 3–62.

Breslow, N. E. 1984. “Extra-Poisson Variation in Log-Linear Models.” *Applied*

Statistics 33 (1): 38. doi:10.2307/2347661.

Bullard, James H, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. 2010. “Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments.” *BMC Bioinformatics* 11 (1) (January): 94. doi:10.1186/1471-2105-11-94.

Burden, Conrad J, Sumaira E Qureshi, and Susan R Wilson. 2014. “Error Estimates for the Analysis of Differential Expression from RNA-Seq Count Data.” *PeerJ* 2 (September 23): e576. doi:10.7717/peerj.576.

Devlin, B., and Kathryn Roeder. 1999. “Genomic Control for Association Studies.” *Biometrics* 55 (4) (December 25): 997–1004. doi:10.1111/j.0006-341X.1999.00997.x.

Dialsingh, Isaac, Stefanie R. Austin, and Naomi S. Altman. 2015. “Estimating the Proportion of True Null Hypotheses When the Statistics Are Discrete.” *Bioinformatics* 31 (14) (July 15): 2303–2309. doi:10.1093/bioinformatics/btv104.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2013. “A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis.” *Briefings in Bioinformatics* 14 (6) (November 1): 671–683. doi:10.1093/bib/bbs046.

Firth, David. 1993. “Bias Reduction of Maximum Likelihood Estimates.” *Biometrika* 80 (1): 27–38. doi:10.1093/biomet/80.1.27.

Furuta, A, L.J Martin, C.-L.G Lin, M Dykes-Hoberg, and J. D. Rothstein. 1997. “Cellular and Synaptic Localization of the Neuronal Glutamate Transporters Excitatory Amino Acid Transporter 3 and 4.” *Neuroscience* 81 (4) (October): 1031–1042. doi:10.1016/S0306-4522(97)00252-2.

Galambos, Janos. 1977. “Bonferroni Inequalities.” *The Annals of Probability* 5 (4): 577–581.

Gao, Xiaoyi, Joshua Starmer, and Eden R Martin. 2008. “A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms.” *Genetic Epidemiology* 32 (4) (May): 361–369. doi:10.1002/gepi.20310.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. “A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models.” *The Annals of Applied Statistics* 2 (4) (December):

1360–1383. doi:10.1214/08-AOAS191.

Han, Fang, and Wei Pan. 2010. “A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants.” *Human Heredity* 70 (1): 42–54. doi:10.1159/000288704.

Heinze, Georg. 2006. “A Comparative Investigation of Methods for Logistic Regression with Separated or Nearly Separated Data.” *Statistics in Medicine* 25 (24) (December 30): 4216–4226. doi:10.1002/sim.2687.

Heinze, Georg, and Michael Schemper. 2002. “A Solution to the Problem of Separation in Logistic Regression.” *Statistics in Medicine* 21 (16) (August 30): 2409–2419. doi:10.1002/sim.1047.

Hendricks, Audrey E, Josée Dupuis, Mark W Logue, Richard H Myers, and Kathryn L Lunetta. 2014. “Correction for Multiple Testing in a Gene Region.” *European Journal of Human Genetics* 22 (3) (March 10): 414–418. doi:10.1038/ejhg.2013.144.

Hommel, G. 1986. “Multiple Test Procedures for Arbitrary Dependence Structures.” *Metrika* 33 (1) (December): 321–336. doi:10.1007/BF01894765.

Jain, N., J. Thatte, T. Braciale, K. Ley, M. O’Connell, and J. K. Lee. 2003. “Local-Pooled-Error Test for Identifying Differentially Expressed Genes with a Small Number of Replicated Microarrays.” *Bioinformatics* 19 (15) (October 12): 1945–1951. doi:10.1093/bioinformatics/btg264.

Jiang, Hui, and Wing Hung Wong. 2009. “Statistical Inferences for Isoform Expression in RNA-Seq.” *Bioinformatics* 25 (8) (April 15): 1026–1032. doi:10.1093/bioinformatics/btp113.

Joe, Harry. 2006. “Generating Random Correlation Matrices Based on Partial Correlations.” *Journal of Multivariate Analysis* 97 (10) (November): 2177–2189. doi:10.1016/j.jmva.2005.05.010.

Labadorf, Adam, Andrew G Hoss, Valentina Lagomarsino, Jeanne C Latourelle, Tiffany C Hadzi, Joli Bregu, Marcy E MacDonald, et al. 2015. “RNA Sequence Analysis of Human Huntington Disease Brain Reveals an Extensive Increase in Inflammatory and Developmental Gene Expression.” Edited by Hiroyoshi Ariga. *PLOS ONE* 10 (12) (December 4): e0143563. doi:10.1371/journal.pone.0143563.

Landau, William Michael, and Peng Liu. 2013. “Dispersion Estimation and Its Effect on Test Performance in RNA-Seq Data Analysis: A Simulation-Based Comparison of Methods.” Edited by Lin Chen. *PLoS ONE* 8 (12) (December

9): e81415. doi:10.1371/journal.pone.0081415.

Lehmann, E. L., and Joseph P. Romano. 2005. “Generalizations of the Familywise Error Rate.” *The Annals of Statistics* 33 (3) (June 1): 1138–1154. doi:10.1214/009053605000000084.

Li, Jun, Daniela M Witten, Iain M Johnstone, and Robert Tibshirani. 2012. “Normalization, Testing, and False Discovery Rate Estimation for RNA-Sequencing Data.” *Biostatistics* 13 (3) (July 1): 523–538. doi:10.1093/biostatistics/kxr031.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12) (December 5): 550. doi:10.1186/s13059-014-0550-8.

MacDonald, Tobey J, Ian F Pollack, Hideho Okada, Soumyaroop Bhattacharya, and James Lyons-Weiler. 2007. “Progression-Associated Genes in Astrocytoma Identified by Novel Microarray Gene Expression Data Reanalysis.” In *Methods in Molecular Biology (Clifton, N.J.)*, 377:203–221. doi:10.1007/978-1-59745-390-5_13.

Marioni, John C, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. 2008. “RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays.” *Genome Research* 18 (9) (July 30): 1509–1517. doi:10.1101/gr.079558.108.

McCarthy, Davis J, Yunshun Chen, and Gordon K Smyth. 2012. “Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation.” *Nucleic Acids Research* 40 (10) (May 1): 4288–4297. doi:10.1093/nar/gks042.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models*. Second. London: Chapman and Hall/CRC Press.

McCullumsmith, R. 2002. “Striatal Excitatory Amino Acid Transporter Transcript Expression in Schizophrenia, Bipolar Disorder, and Major Depressive Disorder.” *Neuropsychopharmacology* 26 (3) (March 17): 368–375. doi:10.1016/S0893-133X(01)00370-0.

Mefford, Joel, and John S Witte. 2012. “The Covariate’s Dilemma.” Edited by Peter M. Visscher. *PLoS Genetics* 8 (11) (November 8): e1003096. doi:10.1371/journal.pgen.1003096.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara

Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7) (July 30): 621–628. doi:10.1038/nmeth.1226.

Nelder, J A, and Y Lee. 1992. "Likelihood, Quasi-Likelihood and Pseudolikelihood: Some Comparisons." *Journal of the Royal Statistical Society. Series B (Methodological)* 54 (1): 273–284.

Nelder, J. A. 2000. "Quasi-Likelihood and Pseudo-Likelihood Are Not the Same Thing." *Journal of Applied Statistics* 27 (8) (November 2): 1007–1011. doi:10.1080/02664760050173328.

Nettleton, Dan, J. T. Gene Hwang, Rico A. Caldo, and Roger P. Wise. 2006. "Estimating the Number of True Null Hypotheses from a Histogram of P Values." *Journal of Agricultural, Biological, and Environmental Statistics* 11 (3) (September): 337–356. doi:10.1198/108571106X129135.

Phipson, Belinda, and Gordon K Smyth. 2010. "Permutation P-Values Should Never Be Zero: Calculating Exact P-Values When Permutations Are Randomly Drawn." *Statistical Applications in Genetics and Molecular Biology* 9 (1) (January 31): Article39. doi:10.2202/1544-6115.1585.

Pickrell, Joseph K, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. 2010. "Understanding Mechanisms Underlying Human Gene Expression Variation with RNA Sequencing." *Nature* 464 (7289) (April 1): 768–772. doi:10.1038/nature08872.

Pirinen, Matti, Peter Donnelly, and Chris C A Spencer. 2012. "Including Known Covariates Can Reduce Power to Detect Genetic Effects in Case-Control Studies." *Nature Genetics* 44 (8) (July 22): 848–851. doi:10.1038/ng.2346.

Pounds, S., and C. Cheng. 2006. "Robust Estimation of the False Discovery Rate." *Bioinformatics* 22 (16) (August 15): 1979–1987. doi:10.1093/bioinformatics/btl328.

Ramsköld, Daniel, Eric T Wang, Christopher B Burge, and Rickard Sandberg. 2009. "An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data." Edited by Lars Juhl Jensen. *PLoS Computational Biology* 5 (12) (December 11): e1000598. doi:10.1371/journal.pcbi.1000598.

Robinson, Laurence D, and Nicholas P Jewell. 1991. "Some Surprising Results about Covariate Adjustment in Logistic Regression Models." *International Statistical Review / Revue Internationale de Statistique* 59 (2) (August): 227. doi:10.2307/1403444.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1) (January 1): 139–140. doi:10.1093/bioinformatics/btp616.

Robinson, Mark D, and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” *Genome Biology* 11 (3) (January): R25. doi:10.1186/gb-2010-11-3-r25.

Robinson, Mark D, and Gordon K Smyth. 2007. “Small-Sample Estimation of Negative Binomial Dispersion, with Applications to SAGE Data.” *Biostatistics* 9 (2) (July 11): 321–332. doi:10.1093/biostatistics/kxm030.

Rocke, David M, Luyao Ruan, Yilun Zhang, J Jared Gossett, Blythe Durbin-Johnson, and Sharon Aviran. 2015. “Excess False Positive Rates in Methods for Differential Gene Expression Analysis Using RNA-Seq Data.” doi:10.1101/020784.

Romano, Joseph P., and Azeem M. Shaikh. 2006. “Stepup Procedures for Control of Generalizations of the Familywise Error Rate.” *The Annals of Statistics* 34 (4) (August 1): 1850–1873. doi:10.1214/009053606000000461.

Si, Yaqing, and Peng Liu. 2013. “An Optimal Test with Maximum Average Power While Controlling FDR with Application to RNA-Seq Data.” *Biometrics* 69 (3) (September 26): 594–605. doi:10.1111/biom.12036.

Soneson, Charlotte, and Mauro Delorenzi. 2013. “A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data.” *BMC Bioinformatics* 14 (1) (January): 91. doi:10.1186/1471-2105-14-91.

Storey, John D, and Robert Tibshirani. 2003. “Statistical Significance for Genomewide Studies.” *Proceedings of the National Academy of Sciences* 100 (16) (August 5): 9440–9445. doi:10.1073/pnas.1530509100.

Storey, John D. 2002. “A Direct Approach to False Discovery Rates.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (3) (August): 479–498. doi:10.1111/1467-9868.00346.

Stuart, Joshua M. 2003. “A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules.” *Science* 302 (5643) (October 10): 249–255. doi:10.1126/science.1087447.

Tarazona, Sonia, F. Garcia-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. 2011. “Differential Expression in RNA-Seq: A Matter of Depth.” *Genome Research* 21 (12) (December 1): 2213–2223.

doi:10.1101/gr.124321.111.

Utal, A.K, A.L Stopka, M Roy, and P.D Coleman. 1998. “PEP-19 Immunohistochemistry Defines the Basal Ganglia and Associated Structures in the Adult Human Brain, and Is Dramatically Reduced in Huntington’s Disease.” *Neuroscience* 86 (4) (June): 1055–1063. doi:10.1016/S0306-4522(98)00130-4.

Vinzenz, Erhardt, and Czado Claudia. 2009. “A Method for Approximately Sampling High-Dimensional Count Variables with Prespecified Pearson Correlation.” http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.147.1847.

Wu, H., Chi Wang, and Zhijin Wu. 2013. “A New Shrinkage Estimator for Dispersion Improves Differential Expression Detection in RNA-Seq Data.” *Biostatistics* 14 (2) (April 1): 232–243. doi:10.1093/biostatistics/kxs033.

Yekutieli, Daniel, and Yoav Benjamini. 2001. “Under Dependency.” *The Annals of Statistics* 29 (4) (August 1): 1165–1188. doi:10.1214/aos/1013699998.

Zaitlen, Noah, Sara Lindström, Bogdan Pasaniuc, Marilyn Cornelis, Giulio Genovese, Samuela Pollack, Anne Barton, et al. 2012. “Informed Conditioning on Clinical Covariates Increases Power in Case-Control Association Studies.” Edited by Peter M. Visscher. *PLoS Genetics* 8 (11) (November 8): e1003032. doi:10.1371/journal.pgen.1003032.

Zorn, C. 2005. “A Solution to Separation in Binary Response Models.” *Political Analysis* 13 (2) (March 1): 157–170. doi:10.1093/pan/mpi009.

CURRICULUM VITAE

Seung Hoan Choi

Department of Biostatistics
Boston University
801 Massachusetts Ave, 3rd Floor
Boston, MA 02118
E-mail: seuchoi@bu.edu

Professional Interests

Statistical Analysis in Genetic and Genomic data, RNA-Sequencing Analysis Methods, Genome-Wide Association Studies, Multiple Testing Correction Methods, Pathway Analysis, Meta-Analysis, Big Data.

Education

Boston University, Ph.D. in Biostatistics (May 2016) Boston, MA
Boston University, M.A. in Biostatistics (May 2011) Boston, MA
The State University of New York at Stony Brook, B.S. in Applied Math & Statistics and Mathematics (Dec. 2008) Stony Brook, NY
Chungnam National University, B.A. in Business Administration (Feb. 2009)
Daejeon, South Korea

Research Experience

2009-Present Research Assistant, Department of Biostatistics, Boston University
2007-2008 Research Assistant, Department of Applied Math & Statistics. The State University of New York at Stony Brook

Teaching Experience

2013-2015 Course Grader (**BS723**: Introduction to statistical Computing, **BS858**: Statistical Genetics I), Department of Biostatistics, Boston University
2012-2014 Teaching Assistant (**Statistical Genetics Section in SIBS**: Summer Institute for Training in Biostatistics), Department of Biostatistics, Boston University
2008 Teaching Assistant (**AMS315**: Data Analysis, **AMS210**: Linear Algebra), Department of Applied Math & Statistics, The State University of New York at Stony Brook

Awards and Honors

2016 Best Poster Presentation, **Cohorts for Heart and Aging Research in Genomic Epidemiology, Houston**
2015 Best Poster Presentation, **Genome Science Institute Research Symposium, Boston University**
2013 CHARGE Rotterdam meeting travel award, **Framingham Heart Study**

2012 CHARGE Reykjavik meeting travel award, **Framingham Heart Study**

2009-Present Graduate Research Assistant Scholarship Program**, Boston University**

2008 Cum Laude, **The State University of New York at Stony Brook**

2008 Undergraduate Research and Creativity Activity Summer Research Fellowship, **The State University of New York at Stony Brook**

Poster Presentations

2016 Evaluation of Logistic Regression Models and Effect of Covariates for Case-Control Study in RNA-Seq Analysis. **Cohorts for Heart and Aging Research in Genomic Epidemiology, Houston, TX**

2015 Evaluation of Logistic Regression Models and Effect of Covariates for Case-Control Study in RNA-Seq Analysis. **Genome Science Institute Research Symposium, Boston, MA**

2014 Six novel loci associated with circulating VEGF levels identified by a meta-analysis genome-wide association study. **The American Society of Human Genetics, San Diego, CA**

2013 Genetic Variants associated with incidence of late-onset Alzheimer's disease in Caucasians. **Alzheimer's Association International Conference, Boston, MA**

2013 Genetic Variants associated with incidence of late-onset Alzheimer's disease in Caucasians. **Cohorts for Heart and Aging Research in Genomic Epidemiology, Rotterdam, Netherland**

2012 Pathway Analysis of Genes Identified by Genome Wide Association Study of Circulating Vascular Endothelial growth factors Levels. **Cohorts for Heart and Aging Research in Genomic Epidemiology, Reykjavik, Iceland**

2011 Pathway Analysis of Genes Identified by Genome Wide Association Study of Circulating Vascular Endothelial growth factors Levels. **Genome Science Institute Research Symposium, Boston, MA**

2009 Growth Mixture Modeling as an Exploratory Analysis Tool in Longitudinal QTL. **Undergraduate Research and Creative Activity, Stony Brook, NY**

Publications

1. **Choi SH**, Ruggiero D, Sorice R, Song C, Nutile T, Vernon Smith A, Concas MP, Traglia M, Barbieri C, Ndiaye NC, Stathopoulou MG, Lagou V, Maestrale GB, Sala C, Debette S, Kovacs P, Lind L, Lamont J, Fitzgerald P, Tönjes A, Gudnason V, Toniolo D, Pirastu M, Bellenguez C, Vasan RS, Ingelsson E, Leutenegger AL, Johnson AD, DeStefano AL, Visvikis-Siest S, Seshadri S, Ciullo M. Six Novel Loci Associated with Circulating VEGF Levels Identified by a Meta-analysis of Genome-Wide Association Studies. PLoS Genet. 2016 Feb 24;12(2):e1005874. doi: 10.1371/journal.pgen.1005874.

2. Jun G, Ibrahim-Verbaas CA, Vronskaya M, Lambert JC, Chung J, Naj AC, Kunkle BW, Wang LS, Bis JC, Bellenguez C, Harold D, Lunetta KL, Destefano AL, Grenier-Boley B, Sims R, Beecham GW, Smith AV, Chouraki V, Hamilton-Nelson KL, Ikram MA, Fievet N, Denning N, Martin ER, Schmidt H, Kamatani Y, Dunstan ML, Valladares O, Laza AR, Zelenika D, Ramirez A, Foroud TM, **Choi SH**, Boland A, Becker T, Kukull WA, van der Lee SJ, Pasquier F, Cruchaga C, Beekly D, Fitzpatrick AL, Hanon O, Gill M, Barber R, Gudnason V, Campion D, Love S, Bennett DA, Amin N, Berr C, Tsolaki M, Buxbaum JD, Lopez OL, Deramecourt V, Fox NC, Cantwell LB, Tárraga L, Dufouil C, Hardy J, Crane PK, Eiriksdottir G, Hannequin D, Clarke R, Evans D, Mosley TH Jr, Letenneur L, Brayne C, Maier W, De Jager P, Emilsson V, Dartigues JF, Hampel H, Kamboh MI, de Bruijn RF, Tzourio C, Pastor P, Larson EB, Rotter JI, O'Donovan MC, Montine TJ, Nalls MA, Mead S, Reiman EM, Jonsson PV, Holmes C, St George-Hyslop PH, Boada M, Passmore P, Wendland JR, Schmidt R, Morgan K, Winslow AR, Powell JF, Carasquillo M, Younkin SG, Jakobsdóttir J, Kauwe JS, Wilhelmsen KC, Rujescu D, Nöthen MM, Hofman A, Jones L; IGAP Consortium, Haines JL, Psaty BM, Van Broeckhoven C, Holmans P, Launer LJ, Mayeux R, Lathrop M, Goate AM, Escott-Price V, Seshadri S, Pericak-Vance MA, Amouyel P, Williams J, van Duijn CM, Schellenberg GD, Farrer LA. A novel Alzheimer disease locus located near the gene encoding tau protein. Mol Psychiatry. 2016 Jan;21(1):108-117. doi: 10.1038/mp.2015.23.

3. Desikan RS, Schork AJ, Wang Y, Witoelar A, Sharma M, McEvoy LK, Holland D, Brewer JB, Chen CH, Thompson WK, Harold D, Williams J, Owen MJ, O'Donovan MC, Pericak-Vance MA, Mayeux R, Haines JL, Farrer LA, Schellenberg GD, Heutink P, Singleton AB, Brice A, Wood NW, Hardy J, Martinez M, **Choi SH**, DeStefano A, Ikram MA, Bis JC, Smith A, Fitzpatrick AL, Launer L, van Duijn C, Seshadri S, Ulstein ID, Aarsland D, Fladby T, Djurovic S, Hyman BT, Snaedal J, Stefansson H, Stefansson K, Gasser T, Andreassen OA, Dale AM. Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus. Mol Psychiatry. 2015 Dec;20(12):1588-95. doi: 10.1038/mp.2015.6.

4. Desikan RS, Schork AJ, Wang Y, Thompson WK, Dehghan A, Ridker PM, Chasman DI, McEvoy LK, Holland D, Chen CH, Karow DS, Brewer JB, Hess CP, Williams J, Sims R, O'Donovan MC, **Choi SH**, Bis JC, Ikram MA, Gudnason V, DeStefano AL, van der Lee SJ, Psaty BM, van Duijn CM, Launer L, Seshadri S, Pericak-Vance MA, Mayeux R, Haines JL, Farrer LA, Hardy J, Ulstein ID, Aarsland D, Fladby T, White LR, Sando SB, Rongve A, Witoelar A, Djurovic S, Hyman BT, Snaedal J, Steinberg S, Stefansson H, Stefansson K, Schellenberg GD, Andreassen OA, Dale AM; Inflammation Working Group and International Genomics of Alzheimer's Disease Project (IGAP) and DemGene

Investigators†. Polygenic Overlap Between C-Reactive Protein, Plasma Lipids, and Alzheimer Disease. Circulation. 2015 Jun 9;131(23):2061-9. doi: 10.1161/CIRCULATIONAHA.115.015489.

5. **International Genomics of Alzheimer's Disease Consortium (IGAP)**. Convergent genetic and expression data implicate immunity in Alzheimer's disease. Alzheimers Dement. 2015 Jun;11(6):658-71. doi: 10.1016/j.jalz.2014.05.1757.

6. Debette S, Ibrahim Verbaas CA, Bressler J, Schuur M, Smith A, Bis JC, Davies G, Wolf C, Gudnason V, Chibnik LB, Yang Q, deStefano AL, de Quervain DJ, Srikanth V, Lahti J, Grabe HJ, Smith JA, Priebe L, Yu L, Karbalai N, Hayward C, Wilson JF, Campbell H, Petrovic K, Fornage M, Chauhan G, Yeo R, Boxall R, Becker J, Stegle O, Mather KA, Chouraki V, Sun Q, Rose LM, Resnick S, Oldmeadow C, Kirin M, Wright AF, Jonsdottir MK, Au R, Becker A, Amin N, Nalls MA, Turner ST, Kardia SL, Oostra B, Windham G, Coker LH, Zhao W, Knopman DS, Heiss G, Griswold ME, Gottesman RF, Vitart V, Hastie ND, Zgaga L, Rudan I, Polasek O, Holliday EG, Schofield P, **Choi SH**, Tanaka T, An Y, Perry RT, Kennedy RE, Sale MM, Wang J, Wadley VG, Liewald DC, Ridker PM, Gow AJ, Pattie A, Starr JM, Porteous D, Liu X, Thomson R, Armstrong NJ, Eiriksdottir G, Assareh AA, Kochan NA, Widen E, Palotie A, Hsieh YC, Eriksson JG, Vogler C, van Swieten JC, Shulman JM, Beiser A, Rotter J, Schmidt CO, Hoffmann W, Nöthen MM, Ferrucci L, Attia J, Uitterlinden AG, Amouyel P, Dartigues JF, Amieva H, Räikkönen K, Garcia M, Wolf PA, Hofman A, Longstreth WT Jr, Psaty BM, Boerwinkle E, DeJager PL, Sachdev PS, Schmidt R, Breteler MM, Teumer A, Lopez OL, Cichon S, Chasman DI, Grodstein F, Müller-Myhsok B, Tzourio C, Papassotiropoulos A, Bennett DA, Ikram MA, Deary IJ, van Duijn CM, Launer L, Fitzpatrick AL, Seshadri S, Mosley TH Jr; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. Genome-wide studies of verbal declarative memory in nondemented older people: the Cohorts for Heart and Aging Research in Genomic Epidemiology consortium. Biol Psychiatry. 2015 Apr 15;77(8):749-63. doi: 10.1016/j.biopsych.2014.08.027.

7. van der Lee SJ, Holstege H, Wong TH, Jakobsdottir J, Bis JC, Chouraki V, van Rooij JG, Grove ML, Smith AV, Amin N, **Choi SH**, Beiser AS, Garcia ME, van IJcken WF, Pijnenburg YA, Louwersheimer E, Brouwer RW, van den Hout MC, Oole E, Eirkisdottir G, Levy D, Rotter JI, Emilsson V, O'Donnell CJ, Aspelund T, Uitterlinden AG, Launer LJ, Hofman A, Boerwinkle E, Psaty BM, DeStefano AL, Scheltens P, Seshadri S, van Swieten JC, Gudnason V, van der Flier WM, Ikram MA, van Duijn CM. PLD3 variants in population studies. Nature. 2015 Apr 2;520(7545):E2-3. doi: 10.1038/nature14038.

8. Nalls MA, Pankratz N, Lill CM, Do CB, Hernandez DG, Saad M, DeStefano AL,

Kara E, Bras J, Sharma M, Schulte C, Keller MF, Arepalli S, Letson C, Edsall C, Stefansson H, Liu X, Pliner H, Lee JH, Cheng R; International Parkinson's Disease Genomics Consortium (IPDGC); Parkinson's Study Group (PSG) Parkinson's Research: The Organized GENetics Initiative (PROGENI); 23andMe; GenePD; NeuroGenetics Research Consortium (NGRC); Hussman Institute of Human Genomics (HIHG); Ashkenazi Jewish Dataset Investigator; **Cohorts for Health and Aging Research in Genetic Epidemiology (CHARGE)**; North American Brain Expression Consortium (NABEC); United Kingdom Brain Expression Consortium (UKBEC); Greek Parkinson's Disease Consortium; Alzheimer Genetic Analysis Group, Ikram MA, Ioannidis JP, Hadjigeorgiou GM, Bis JC, Martinez M, Perlmutter JS, Goate A, Marder K, Fiske B, Sutherland M, Xiromerisiou G, Myers RH, Clark LN, Stefansson K, Hardy JA, Heutink P, Chen H, Wood NW, Houlden H, Payami H, Brice A, Scott WK, Gasser T, Bertram L, Eriksson N, Foroud T, Singleton AB. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nat Genet. 2014 Sep;46(9):989-93. doi: 10.1038/ng.3043.

9. Ibrahim-Verbaas CA, Fornage M, Bis JC, **Choi SH**, Psaty BM, Meigs JB, Rao M, Nalls M, Fontes JD, O'Donnell CJ, Kathiresan S, Ehret GB, Fox CS, Malik R, Dichgans M, Schmidt H, Lahti J, Heckbert SR, Lumley T, Rice K, Rotter JI, Taylor KD, Folsom AR, Boerwinkle E, Rosamond WD, Shahar E, Gottesman RF, Koudstaal PJ, Amin N, Wieberdink RG, Dehghan A, Hofman A, Uitterlinden AG, Destefano AL, Debette S, Xue L, Beiser A, Wolf PA, Decarli C, Ikram MA, Seshadri S, Mosley TH Jr, Longstreth WT Jr, van Duijn CM, Launer LJ. Predicting stroke through genetic risk functions: the CHARGE Risk Score Project. Stroke. 2014 Feb;45(2):403-12. doi: 10.1161/STROKEAHA.113.003044.

10. Weinstein G, Beiser AS, **Choi SH**, Preis SR, Chen TC, Vorgas D, Au R, Pikula A, Wolf PA, DeStefano AL, Vasan RS, Seshadri S. Serum brain-derived neurotrophic factor and the risk for dementia: the Framingham Heart Study. JAMA Neurol. 2014 Jan;71(1):55-61. doi: 10.1001/jamaneurol.2013.4781.

11. Escott-Price V, Bellenguez C, Wang LS, **Choi SH**, Harold D, Jones L, Holmans P, Gerrish A, Vedernikov A, Richards A, DeStefano AL, Lambert JC, Ibrahim-Verbaas CA, Naj AC, Sims R, Jun G, Bis JC, Beecham GW, Grenier-Boley B, Russo G, Thornton-Wells TA, Denning N, Smith AV, Chouraki V, Thomas C, Ikram MA, Zelenika D, Vardarajan BN, Kamatani Y, Lin CF, Schmidt H, Kunkle B, Dunstan ML, Vronskaya M; United Kingdom Brain Expression Consortium, Johnson AD, Ruiz A, Bihoreau MT, Reitz C, Pasquier F, Hollingworth P, Hanon O, Fitzpatrick AL, Buxbaum JD, Campion D, Crane PK, Baldwin C, Becker T, Gudnason V, Cruchaga C, Craig D, Amin N, Berr C,

Lopez OL, De Jager PL, Deramecourt V, Johnston JA, Evans D, Lovestone S, Letenneur L, Hernández I, Rubinsztein DC, Eiriksdottir G, Sleegers K, Goate AM, Fiévet N, Huentelman MJ, Gill M, Brown K, Kamboh MI, Keller L, Barberger-Gateau P, McGuinness B, Larson EB, Myers AJ, Dufouil C, Todd S, Wallon D, Love S, Rogaeva E, Gallacher J, George-Hyslop PS, Clarimon J, Lleo A, Bayer A, Tsuang DW, Yu L, Tsolaki M, Bossù P, Spalletta G, Proitsi P, Collinge J, Sorbi S, Garcia FS, Fox NC, Hardy J, Naranjo MC, Bosco P, Clarke R, Brayne C, Galimberti D, Scarpini E, Bonuccelli U, Mancuso M, Siciliano G, Moebus S, Mecocci P, Zompo MD, Maier W, Hampel H, Pilotto A, Frank-García A, Panza F, Solfrizzi V, Caffarra P, Nacmias B, Perry W, Mayhaus M, Lannfelt L, Hakonarson H, Pichler S, Carrasquillo MM, Ingelsson M, Beekly D, Alvarez V, Zou F, Valladares O, Younkin SG, Coto E, Hamilton-Nelson KL, Gu W, Razquin C, Pastor P, Mateo I, Owen MJ, Faber KM, Jonsson PV, Combarros O, O'Donovan MC, Cantwell LB, Soininen H, Blacker D, Mead S, Mosley TH Jr, Bennett DA, Harris TB, Fratiglioni L, Holmes C, de Bruijn RF, Passmore P, Montine TJ, Bettens K, Rotter JI, Brice A, Morgan K, Foroud TM, Kukull WA, Hannequin D, Powell JF, Nalls MA, Ritchie K, Lunetta KL, Kauwe JS, Boerwinkle E, Riemenschneider M, Boada M, Hiltunen M, Martin ER, Schmidt R, Rujescu D, Dartigues JF, Mayeux R, Tzourio C, Hofman A, Nöthen MM, Graff C, Psaty BM, Haines JL, Lathrop M, Pericak-Vance MA, Launer LJ, Van Broeckhoven C, Farrer LA, van Duijn CM, Ramirez A, Seshadri S, Schellenberg GD, Amouyel P, Williams J; Cardiovascular Health Study (CHS). Gene-wide analysis detects two new susceptibility genes for Alzheimer's disease. PLoS One. 2014 Jun 12;9(6):e94661. doi: 10.1371/journal.pone.0094661

12. Bis JC, DeStefano A, Liu X, Brody JA, **Choi SH**, Verhaaren BF, Debette S, Ikram MA, Shahar E, Butler KR Jr, Gottesman RF, Muzny D, Kovar CL, Psaty BM, Hofman A, Lumley T, Gupta M, Wolf PA, van Duijn C, Gibbs RA, Mosley TH, Longstreth WT Jr, Boerwinkle E, Seshadri S, Fornage M. Associations of NINJ2 sequence variants with incident ischemic stroke in the Cohorts for Heart and Aging in Genomic Epidemiology (CHARGE) consortium. PLoS One. 2014 Jun 24;9(6):e99798. doi: 10.1371/journal.pone.0099798

13. Chen H, **Choi SH**, Hong J, Lu C, Milton JN, Allard C, Lacey SM, Lin H, Dupuis J. Rare genetic variant analysis on blood pressure in related samples. BMC Proc. 2014 Jun 17;8(Suppl 1):S35. doi: 10.1186/1753-6561-8-S1-S35.

14. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStafano AL, Bis JC, Beecham GW, Grenier-Boley B, Russo G, Thorton-Wells TA, Jones N, Smith AV, Chouraki V, Thomas C, Ikram MA, Zelenika D, Vardarajan BN, Kamatani Y, Lin CF, Gerrish A, Schmidt H, Kunkle B, Dunstan ML, Ruiz A, Bihoreau MT, **Choi SH**, Reitz C, Pasquier F, Cruchaga C, Craig

D, Amin N, Berr C, Lopez OL, De Jager PL, Deramecourt V, Johnston JA, Evans D, Lovestone S, Letenneur L, Morón FJ, Rubinsztein DC, Eiriksdottir G, Sleegers K, Goate AM, Fiévet N, Huentelman MW, Gill M, Brown K, Kamboh MI, Keller L, Barberger-Gateau P, McGuiness B, Larson EB, Green R, Myers AJ, Dufouil C, Todd S, Wallon D, Love S, Rogaeva E, Gallacher J, St George-Hyslop P, Clarimon J, Lleo A, Bayer A, Tsuang DW, Yu L, Tsolaki M, Bossù P, Spalletta G, Proitsi P, Collinge J, Sorbi S, Sanchez-Garcia F, Fox NC, Hardy J, Deniz Naranjo MC, Bosco P, Clarke R, Brayne C, Galimberti D, Mancuso M, Matthews F; European Alzheimer's Disease Initiative (EADI); Genetic and Environmental Risk in Alzheimer's Disease; Alzheimer's Disease Genetic Consortium; Cohorts for Heart and Aging Research in Genomic Epidemiology, Moebus S, Mecocci P, Del Zompo M, Maier W, Hampel H, Pilotto A, Bullido M, Panza F, Caffarra P, Nacmias B, Gilbert JR, Mayhaus M, Lannefelt L, Hakonarson H, Pichler S, Carrasquillo MM, Ingelsson M, Beekly D, Alvarez V, Zou F, Valladares O, Younkin SG, Coto E, Hamilton-Nelson KL, Gu W, Razquin C, Pastor P, Mateo I, Owen MJ, Faber KM, Jonsson PV, Combarros O, O'Donovan MC, Cantwell LB, Soininen H, Blacker D, Mead S, Mosley TH Jr, Bennett DA, Harris TB, Fratiglioni L, Holmes C, de Bruijn RF, Passmore P, Montine TJ, Bettens K, Rotter JI, Brice A, Morgan K, Foroud TM, Kukull WA, Hannequin D, Powell JF, Nalls MA, Ritchie K, Lunetta KL, Kauwe JS, Boerwinkle E, Riemenschneider M, Boada M, Hiltuenen M, Martin ER, Schmidt R, Rujescu D, Wang LS, Dartigues JF, Mayeux R, Tzourio C, Hofman A, Nöthen MM, Graff C, Psaty BM, Jones L, Haines JL, Holmans PA, Lathrop M, Pericak-Vance MA, Launer LJ, Farrer LA, van Duijn CM, Van Broeckhoven C, Moskvina V, Seshadri S, Williams J, Schellenberg GD, Amouyel P. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013 Dec;45(12):1452-8. doi: 10.1038/ng.2802.

15. Schilling S, DeStefano AL, Sachdev PS, **Choi SH**, Mather KA, DeCarli CD, Wen W, Høgh P, Raz N, Au R, Beiser A, Wolf PA, Romero JR, Zhu YC, Lunetta KL, Farrer L, Dufouil C, Kuller LH, Mazoyer B, Seshadri S, Tzourio C, Debette S. APOE genotype and MRI markers of cerebrovascular disease: systematic review and meta-analysis. Neurology. 2013 Jul 16;81(3):292-300. doi: 10.1212/WNL.0b013e31829bfda4.

16. Stein JL, Medland SE, Vasquez AA, Hibar DP, Senstad RE, Winkler AM, Toro R, Appel K, Bartecek R, Bergmann Ø, Bernard M, Brown AA, Cannon DM, Chakravarty MM, Christoforou A, Domin M, Grimm O, Hollinshead M, Holmes AJ, Homuth G, Hottenga JJ, Langan C, Lopez LM, Hansell NK, Hwang KS, Kim S, Laje G, Lee PH, Liu X, Loth E, Lourdusamy A, Mattingsdal M, Mohnke S, Maniega SM, Nho K, Nugent AC, O'Brien C, Papmeyer M, Pütz B, Ramasamy A, Rasmussen J, Rijpkema M, Risacher SL, Roddey JC, Rose EJ, Ryten M, Shen L, Sprooten E, Strengman E, Teumer A, Trabzuni D, Turner J, van Eijk

K, van Erp TG, van Tol MJ, Wittfeld K, Wolf C, Woudstra S, Aleman A, Alhusaini S, Almasy L, Binder EB, Brohawn DG, Cantor RM, Carless MA, Corvin A, Czisch M, Curran JE, Davies G, de Almeida MA, Delanty N, Depondt C, Duggirala R, Dyer TD, Erk S, Fagerness J, Fox PT, Freimer NB, Gill M, Göring HH, Hagler DJ, Hoehn D, Holsboer F, Hoogman M, Hosten N, Jahanshad N, Johnson MP, Kasperaviciute D, Kent JW Jr, Kochunov P, Lancaster JL, Lawrie SM, Liewald DC, Mandl R, Matarin M, Mattheisen M, Meisenzahl E, Melle I, Moses EK, Mühleisen TW, Nauck M, Nöthen MM, Olvera RL, Pandolfo M, Pike GB, Puls R, Reinvang I, Rentería ME, Rietschel M, Roffman JL, Royle NA, Rujescu D, Savitz J, Schnack HG, Schnell K, Seiferth N, Smith C, Steen VM, Valdés Hernández MC, Van den Heuvel M, van der Wee NJ, Van Haren NE, Veltman JA, Völzke H, Walker R, Westlye LT, Whelan CD, Agartz I, Boomsma DI, Cavalleri GL, Dale AM, Djurovic S, Drevets WC, Hagoort P, Hall J, Heinz A, Jack CR Jr, Foroud TM, Le Hellard S, Macciardi F, Montgomery GW, Poline JB, Porteous DJ, Sisodiya SM, Starr JM, Sussmann J, Toga AW, Veltman DJ, Walter H, Weiner MW; Alzheimer's Disease Neuroimaging Initiative; EPIGEN Consortium; IMAGEN Consortium; Saguenay Youth Study Group, Bis JC, Ikram MA, Smith AV, Gudnason V, Tzourio C, Vernooij MW, Launer LJ, DeCarli C, Seshadri S; **Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium**, Andreassen OA, Apostolova LG, Bastin ME, Blangero J, Brunner HG, Buckner RL, Cichon S, Coppola G, de Zubicaray GI, Deary IJ, Donohoe G, de Geus EJ, Espeseth T, Fernández G, Glahn DC, Grabe HJ, Hardy J, Hulshoff Pol HE, Jenkinson M, Kahn RS, McDonald C, McIntosh AM, McMahon FJ, McMahon KL, Meyer-Lindenberg A, Morris DW, Müller-Myhsok B, Nichols TE, Ophoff RA, Paus T, Pausova Z, Penninx BW, Potkin SG, Sämann PG, Saykin AJ, Schumann G, Smoller JW, Wardlaw JM, Weale ME, Martin NG, Franke B, Wright MJ, Thompson PM; Enhancing Neuro Imaging Genetics through Meta-Analysis Consortium. Identification of common variants associated with human hippocampal and intracranial volumes. Nat Genet. 2012 Apr 15;44(5):552-61. doi: 10.1038/ng.2250.

17. Bis JC, DeCarli C, Smith AV, van der Lijn F, Crivello F, Fornage M, Debette S, Shulman JM, Schmidt H, Srikanth V, Schuur M, Yu L, **Choi SH**, Sigurdsson S, Verhaaren BF, DeStefano AL, Lambert JC, Jack CR Jr, Struchalin M, Stankovich J, Ibrahim-Verbaas CA, Fleischman D, Zijdenbos A, den Heijer T, Mazoyer B, Coker LH, Enzinger C, Danoy P, Amin N, Arfanakis K, van Buchem MA, de Bruijn RF, Beiser A, Dufouil C, Huang J, Cavalieri M, Thomson R, Niessen WJ, Chibnik LB, Gislason GK, Hofman A, Pikula A, Amouyel P, Freeman KB, Phan TG, Oostra BA, Stein JL, Medland SE, Vasquez AA, Hibar DP, Wright MJ, Franke B, Martin NG, Thompson PM; Enhancing Neuro Imaging Genetics through Meta-Analysis Consortium, Nalls MA, Uitterlinden AG, Au R, Elbaz A, Beare RJ, van Swieten JC, Lopez OL, Harris TB, Chouraki V, Breteler MM, De Jager PL, Becker JT, Vernooij MW, Knopman D, Fazekas F, Wolf PA, van der Lugt A, Gudnason V, Longstreth WT Jr, Brown MA, Bennett

DA, van Duijn CM, Mosley TH, Schmidt R, Tzourio C, Launer LJ, Ikram MA, Seshadri S; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. Common variants at 12q14 and 12q24 are associated with hippocampal volume. Nat Genet. 2012 Apr 15;44(5):545-51. doi: 10.1038/ng.2237.

18. **Choi SH**, Liu C, Dupuis J, Logue MW, Jun G. Using linkage analysis of large pedigrees to guide association analyses. BMC Proc. 2011 Nov 29;5 Suppl 9:S79. doi: 10.1186/1753-6561-5-S9-S79.

19. Debette S, Visvikis-Siest S, Chen MH, Ndiaye NC, Song C, Destefano A, Safa R, Azimi Nezhad M, Sawyer D, Marteau JB, Xanthakis V, Siest G, Sullivan L, Pfister M, Smith H, **Choi SH**, Lamont J, Lind L, Yang Q, Fitzgerald P, Ingelsson E, Vasan RS, Seshadri S. Identification of cis- and trans-acting genetic variants explaining up to half the variation in circulating vascular endothelial growth factor levels. Circ Res. 2011 Aug 19;109(5):554-63. doi: 10.1161/CIRCRESAHA.111.243790.

20. Chang SW, **Choi SH**, Li K, Fleur RS, Huang C, Shen T, Ahn K, Gordon D, Kim W, Wu R, Mendell NR, Finch SJ. Growth mixture modeling as an exploratory analysis tool in longitudinal quantitative trait loci analysis. BMC Proc. 2009 Dec 15;3 Suppl 7:S112.

21. Huang C, Li K, Fleur RS, Chang SW, **Choi SH**, Shen T, Shin SY, Finch SJ, Mendell NR. Family-based analysis of a myocardial infarction endophenotype: comparison of sampling designs. BMC Proc. 2009 Dec 15;3 Suppl 7:S120.