Evaluation of Logistic Regression Models and Effect of Covariates for Case-Control Study in RNA-seq Analysis

Seung-Hoan Choi*[1], Adam Labadorf[2, 3], Richard H Myers[2, 4], Kathryn Lunetta[1], Josée Dupuis[1,3], Anita L DeStefano**[1,4]

[1]Department of Biostatistics, School of Public Health, Boston University, Boston, MA, USA

[2]Department of Neurology, Boston University School of Medicine, Boston, MA, USA

[3]Bioinformatics Program, Boston University, Boston, MA, USA

[4]Genome Science Institute, Boston University School of Medicine, Boston, MA, USA

Introduction

Recent Next Generation Sequencing (NGS) methods provide an unbiased count of RNA molecules in the form of RNA-sequencing (RNA-seq) reads, yielding discrete, often highly non-normally distributed gene expression measurements. 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. Because many RNA-seq studies are designed to compare cases and controls, we explore logistic regression as an alternative method, where we model disease status as a function of RNA-seq reads. 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.

Methods

To evaluate the appropriateness of distinct analysis methods for RNA-seq data, we explored NB regression and logistic regression by applying Classical Logistic (CL), Bayes Logistic (BL), and Firth Logistic (FL) regression approaches. To remedy inaccurate test statistic asymptotic distributions yielding incorrect type-I error rate control, we approximate the distribution of the test statistic under the null hypothesis of no association using the data adaptive method suggested by Han and Pan (2010). We also examined the performance of various covariate models with NB regression. We used both simulated data sets and a real Huntington's disease (HD) RNA-seq data set to assess different methods and scenarios.

Results

When the sample size is small or the expression of a gene are highly dispersed the NB regression shows inflated type-I error but the CL and BL regressions are conservative. Large sample size and low dispersion generally make type-I error rates of all methods close to nominal alpha levels of 0.05 and 0.01. The FL regression performs well or is somewhat

conservative at our alpha levels. For all regression methods, power is enhanced by low dispersion, large mean expression value in controls, and large $\log_2$fold-change. The NB, BL, and FL regressions gain increased power with large sample size, large $\log_2$fold-change, and low dispersion.

Increasing the number of non-predictive covariates inflates type-I error rates in the NB regression, when the sample size is small. The change of effect size of a covariate does not affect type-I error rates. The power of the NB regression is decreased by the increase of effect size of a covariate and the number of NP covariates, when the sample size is small. Both larger sample sizes and larger effect sizes increase the power of the NB regression.

Conclusion

We recommend implementing the data adaptive method to evaluate the significance of the analysis procedure of RNA-seq studies because Type-I error rates vary by regression methods and genes. Considering the computational burden of the data adaptive method, the FL regression is the best option for controlling Type-I error with comparable power to the NB regression if the sample size is not too small. A parsimonious model, limiting the number of covariates, is necessary to obtain robust results in the NB regression.