Multiple Hypothesis Testing in RNA Sequencing Gene Isoform Expression Analysis

Bo Li[*]

Abstract
In this article, we overview multiple hypothesis testing procedures in detecting differentially expressed gene isoforms based on generalized linear models. We apply these methods to a real RNA sequencing data for illustration.

Key Words: RNA Sequencing Data; Multiple Hypothesis Tests

1. Introduction

In genome studies, researchers often conduct microarray or RNA sequencing experiments to identify differentially expressed genes. A major concern in high-dimensional gene expression analysis is to control the family-wise error rate (FWER). Popular single-step multiple hypothesis testing methods include Bonferroni method and Scheff$\acute{e}$'s method. Holm's step-down method is known to provide more powerful tests as compared with Bonferroni method. Benjamini and Hochberg (BH) step-up method is often used to control the false discovery rate (FDR). As a reminder, FDR is the expected value of the ratio of the number of false positives to all misexpressed genes detected. It is defined as 0 if no misexpressed genes were found. Note that if a method strongly controls FWER then it controls FDR, Dudoit et al (2003). Multiple comparison procedures in microarray gene expression analysis have been well developed, Dudoit et al (2002), Dudoit et al (2003), Hsu et al (2006), and Li and Mansouri (2016). Multiple hypothesis testing methods based on normal theory are implemented in package edgeR (Chen et al, 2019) in detecting differentially expressed genes for RNA sequencing data. For small-sample experiments, Li et al (2012) proposed the significance analysis of sequencing data based on permutation to control the FDR. We provide an overview of multiple hypothesis testing methods based on normal theory of Chen et al (2014) and the resampling method of Li et al (2012) in this article. A real example is used to illustrate the application.

2. Multiple Hypothesis Testing based on Normal Theory

In this section, we review multiple hypothesis testing methods based on a negative binomial generalized linear model. For gene $l,\ l = 1, \cdots, g$ let Y_{lij} be normalized observation from $i - th$ treatment and $j - th$ block, $i = 1, 2;\ j = 1, \cdots, b$. Chen at el (2014) assume $Y_{lij}|\phi$ follows negative binomial distribution with mean μ_{lij} and variance $\mu_{lij} + \phi\mu_{lij}^2$ where ϕ is the overall dispersion parameter. Alternatively, gene-wise dispersion parameter $\phi_l,\ l = 1, \cdots, g$ may be used if exploratory analysis, for instance, BCV plot of Chen et al (2019), shows "apparent" heterogeneity among the value of the biological coefficient of variation (BCV), see Chen et al (2014) for details. We assume normalized observations follow a per gene generalized linear model that

$$log(\mu_{lij}) = G_l + T_{li} + B_{lj} \tag{2.1}$$

[*] Department of Mathematical Sciences, The Citadel, The Military College of South Carolina, Charleston, SC, 29409. Email: bli@citadel.edu

where G_l is the overall mean expression from gene l, $l = 1, \cdots, g$; T_{li} is the $i - th$ treatment effect on $l - th$ gene with $T_{l1} + T_{l2} = 0$ for all l; B_{lj} is the $j - th$ block effect on $l - th$ gene with $\sum_j B_{lj} = 0$ for all l. We let $\boldsymbol{Y}_l$ be the vector of observations from gene l, for all $l = 1, \cdots, g$. Write $\boldsymbol{T}_l = (T_{l1}, T_{l2})'$ and $\boldsymbol{B}_l = (B_{l1}, \cdots, B_{lb})'$.

To identify differentially expressed genes, we test a sequence of hypotheses that

$$H_{0_l} : T_{l1} - T_{l2} = 0 \ \ vs. \ \ H_{1_l} : T_{l1} - T_{l2} \neq 0 \tag{2.2}$$

for $l = 1, \cdots, g$.

The reduced model under null hypothesis H_{0_l} can be written as

$$log(\mu_{lij}) = G_l + B_{lj} \tag{2.3}$$

for $l = 1, \cdots, g$.

Let $l(\phi; \boldsymbol{Y}_l, \widehat{\boldsymbol{T}}_l, \widehat{\boldsymbol{B}}_l)$ and $l(\phi; \boldsymbol{Y}_l, \widehat{\boldsymbol{B}}_l)$ be the log-likelihood function under the full model (2.1) and the reduced model (2.3) respectively. $\widehat{\boldsymbol{T}}_l$ and $\widehat{\boldsymbol{B}}_l$ are maximum likelihood estimation of the parameters. The likelihood ratio test statistic (LRT) of Chen et al (2014) takes the form that

$$LRT_l = 2[l(\phi; \boldsymbol{Y}_l, \widehat{\boldsymbol{T}}_l, \widehat{\boldsymbol{B}}_l) - l(\phi; \boldsymbol{Y}_l, \widehat{\boldsymbol{B}}_l)] \tag{2.4}$$

for $l = 1, \cdots, g$.

The method of Chen et al (2014) approximates $p - value$ by comparing the likelihood ratio test statistic value of (2.4) with $Chi - Square$ distribution with degree of freedom $df = b - 1$. The resulting $p - value$ are adjusted using Bonferroni method, Holm's step-down method, or BH step-up method in edgeR. As compared with Holm's step-down method and BH step-up method, Bonferroni method is known to be conservative in gene expression analysis, Dudoit et al (2003). We focus on Holm's step-down method and BH step-up method in this article.

To proceed, denote the ordered $p - value$ by

$$p_{r_1} \leq p_{r_2} \cdots p_{r_{l-1}} \leq p_{r_l} \leq p_{r_{l+1}} \cdots \leq p_{r_g}$$

which are associated to hypotheses $H_{0_{r_1}}, H_{0_{r_2}}, \cdots, H_{0_{r_{l-1}}}, H_{0_{r_l}}, H_{0_{r_{l+1}}}, \cdots, H_{0_{r_g}}$.

Holm's step-down method adjusts $l - th, \ l = 1, \cdots, g$ ordered $p - value$ by

$$p_{r_l}^{Holm} = \max_{k=1,\cdots,l} \{\min[(g - k + 1)p_{r_k}, 1]\} \tag{2.5}$$

Holm's method controls the family-wise error rate. Researchers often attempt to control the false discovery rate in RNA sequencing gene expression analysis. BH step-up method provides the control of FDR and it calibrates the ordered $p - value$ in the form that

$$p_{r_l}^{BH} = \min_{k=l,\cdots,g} \{\min[(g - k + 1)p_{r_k}, 1]\} \tag{2.6}$$

For details, see page 79 and 80 of Dudoit et al (2003).

3. Significance Analysis of Sequencing Data

It is required by BH step-up method that $p - value$ are accurate. Since sample size of RNA sequencing data is often small, the testing result of BH method in section 2 is questionable. Li et al (2012) proposed the method, namely the significance analysis of sequencing data, based on permutation under complete null hypothesis whose elementary hypothesis is in

(2.2). In brief, power transformed observations Y_{lij} are assumed to follow $Poisson(\mu_{li})$ for all $j = 1, \cdots, b$ which are fit to a log-linear model that

$$log(\mu_{li}) = log(d_i) + G_l + T_{li} \tag{3.7}$$

where d_i denotes the normalized library size for the readings from $i-th$ treatment, $i = 1, 2$. Iterative normalization procedure of Li et al (2012) is used to approximate the library size $d_i,\ i = 1, 2$. To test a sequence of hypotheses in (2.2) in association to the model in (3.7), the score test statistic of Li et al (2012) for gene $l,\ l = 1, \cdots, g$ is given by

$$S_l = \frac{(Y_{l1\cdot} - Y_{l1\cdot}^{(0)})^2}{Y_{l1\cdot}^{(0)}} + \frac{(Y_{l2\cdot} - Y_{l2\cdot}^{(0)})^2}{Y_{l2\cdot}^{(0)}} \tag{3.8}$$

where $Y_{li\cdot} = \sum_{j=1}^{b} Y_{lij}$ and $Y_{li\cdot}^{(0)} = b\widehat{\mu}_{li}^{(0)},\ i = 1, 2$. Note that $\widehat{\mu}_{li}^{(0)}$ is computed using maximum likelihood estimation based on the reduced model in section 3.2 of Li et al (2012).

The FDR of Li et al (2012) is estimated by the following steps.

1) Resample observations from each gene without replacement B times. Let $S_1^{(r)}, \cdots,$ $S_g^{(r)}$ be the statistic value computed based on $r-th$ permutation data set using the equation in (3.8), $r = 1, \cdots, B$.

2) Let $\widehat{R}_c$ be the total number of genes rejected by comparing the statistic value in (3.8) with a cut point c. Let $\widehat{V}_c$ be the bootstrap mean of “false positives” among all rejected genes. For the set of “false positives”, see the approximation method in section 4 of Li et al (2012).

The false discovery rate is computed by $FDR = \frac{\widehat{V}_c}{\widehat{R}_c}$. We plot a sequence of cut points c versus their corresponding FDR’s. Set the nominal level of the FDR as α and denote c_α the associated cut point. A gene is detected as differentially expressed if the test statistic value in (3.8) is greater than c_α.

4. Example

We apply the multiplicity adjustment methods in section 2 and 3 to a real RNA sequencing data. The data is from Tuch et al (2010). RNA samples are from 3 patients (block factor) and normal and tumor cells (treatment factor) of each patient are used for extraction. The RNA samples are sequenced in 6 sequencing devices. The resulting measurements are normalized using upper-quantile normalization procedure of Bullard et al (2010), where the $0.75-th$ quantile of each library is used to approximate the library size. The normalized observations are fit to the generalized linear model in (2.1). To test hypotheses in (2.2), the asymptotic $p-value$ are adjusted using Bonferroni method, Holm’s step-down method, and BH step-up method in section 2 respectively. Moreover, we apply the significance analysis of sequencing data of Li et al (2012) in section 3 to detect differentially expressed genes using package PoissonSeq of Li (2012). The results are summarized in Figure 1. The Venn diagram shows that BH step-up method identifies more differentially expressed genes than step-down and single-step methods. Permutation based method of Li et al (2012) uniquely detects 73 misexpressed genes. In section 5 of Li et al (2012), it shows that the significance analysis of sequencing data provides an estimation of the FDR closer to the “ture” level as compared with BH step-up method. Since the set of misexpressed genes (the list of misexpressed genes is available on request from the author (bli@citadel.edu)) varies from one method to another, further investigation is necessary.

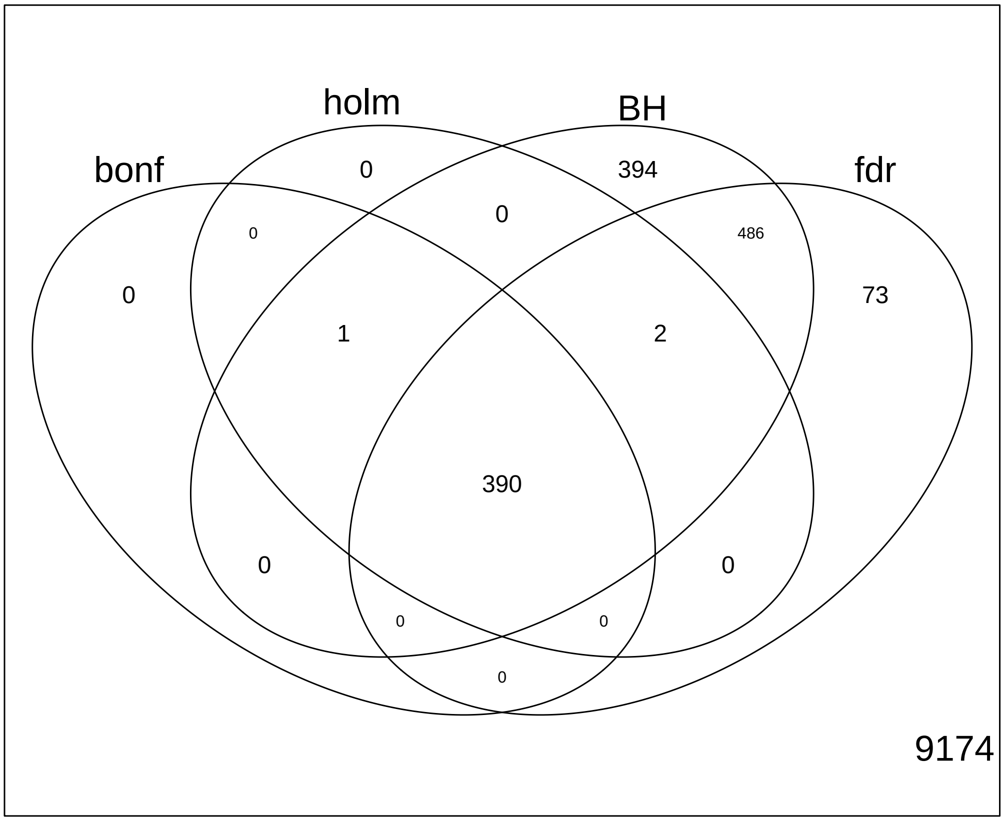

Figure 1: Differentially Expressed Genes. (Note: “bonf”, “holm”, and “BH” denote Bonferroni method, Holm’s step-down method, and BH step-up method in section 2 respectively. “fdr” denotes the significance analysis of sequencing data of Li et el (2012) in section 3. The permutation size $B = 100$ to estimate the FDR. The number listed inside and outside a circle means the number of differentially and not differentially expressed genes respectively.)

REFERENCES

Bullard J. H, Purdom E., Hansen K D., Dudoit S. (2010), “Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments,” *BMC bioinformatics*, 11, 94.

Chen Y.S., Lun A.T.L., Smyth G.K. (2014), “Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR,” in *Statistical Analysis of Next Generation Sequencing Data*, eds. Somnath Datta and Daniel S Nettleton, New York: Springer.

Chen Y., McCarthy D., Robinson M., Smyth G.K. (2019), “edgeR: differential expression analysis of digital gene expression data User’s Guide,” https://bioconductor.org/packages/release/bioc/manuals/edgeR/man/edgeR.pdf.

Dudoit S., Yang Y.H., Callow M.J., Speed T.P. (2001), “Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments,” *Statistica Sinica*, 12, 111–139.

Dudoit S., Shaffer J.P., Boldrick J.C. (2003), “Multiple Hypothesis Testing in Microarray Experiments,” *Statistical Science*,18(1), 71–103.

Hsu J.C., Chang J.Y., Wang T. (2006), “Simultaneous confidence intervals for differential gene expressions,” *Journal of Statistical Planning and Inference*, 136(7), 2182–2196.

Li B., Mansouri H.G. (2016), “Simultaneous Rank Tests for Detecting Differentially Expressed Genes,” *Journal of Statistical Computation and Simulation*, 86(5), 959–972.

Li J., Witten D.M., Johnstone I.M., Tibshirani R. (2012), “Normalization, testing, and false discovery rate estimation for RNA-sequencing data,” *JBiostatistics*, 13(3), 523–538.

Li J. (2012), “PoissonSeq: Significance analysis of sequencing data based on a Poisson log linear model,” https://cran.r-project.org/web/packages/PoissonSeq/PoissonSeq.pdf.

Tuch B.B., Laborde R.R., Xu X., Gu J., Chung C.B., Monighetti C.K., Stanley S.J., Olsen K.D., Kasperbauer J.L., Moore E.J., Broomer A.J., Tan R., Brzoska P.M., Muller M.W., Siddiqui A.S., Asmann Y.W., Sun Y., Kuersten S., Barker M.A., Vega F.M.D.L., Smith D.I. (2010), “Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations,” *PLoS ONE*, 5, e9317.