Identifying the differentially expressed genes with RNA-Seq data

Hung-Ting Lu, Huey-Miin Hsueh
Department of Statistics, National Chengchi University, Taipei, Taiwan
Corresponding author:Huey-Miin Hsueh, email:hsueh@nccu.edu.tw

April 22, 2013

Abstract

Recently, the RNA-Seq experiment is developed for a high-throughput DNA sequencing method for mapping and quantifying the transcriptomes. The gene expression level obtained from a RNA-Seq experiment is of the count data type and is often fitted by a Negative-Binomial distribution to account for over-dispersion. To find the differentially expressed genes with a binary phenotypic response, we aim to develop a statistical test for comparing the means of two Negative-Binomial distributions. A Wald test statistic based on the pseudo maximum likelihood estimators is proposed. A numerical study is performed for justification of the proposed test. The applicability of the proposed method is demonstrated via the data analysis of two real example data sets.

Keywords: Differentially expressed genes, negative-binomial, overdispersion, pseudo maximum likelihood estimator, RNA-Seq experiment.

1 Introduction

To quantify the transcripts in a cell, the hybridization-based approaches, such as microarrays, have become prominent due to the fact that they are high throughput and cost less. On the other hand, although the traditional sequence-based approach can directly provide the cDNA sequence, it is of limited use because it is relatively less effective and expensive. Recently, the RNA-Seq experiment is developed for a high-throughput DNA sequencing method for mapping and quantifying the transcriptomes. The method improves the existing microarray approaches in terms of providing more accurate signals, being not limited to existing genomic sequence and requiring less RNA sample. The RNA-Seq is believed to replace microarrays when the sequencing cost reduces. Please see Wang, Gerstein and Snyder (2009) for a thorough description of the experiment.

This study aims to develop a statistical testing procedure to determine whether a gene is differentially expressed in different experimental/phenotypic conditions. For simplicity, we consider a binary phenotypic response variable. It's known that the gene expression level obtained from a RNA-Seq experiment, differing with that from a microarray experiment, is the number of repeated reads and hence is of the count data type. The researchers assume a poisson population, or a negative-binomial population to account for over-dispersion, for the expression level of an individual gene. See Robinson and Smyth (2007, 2008), Robinson et al. (2010), Li et al. (2012). Due to having simpler formulations, most of the existing methods consider the likelihood ratio test or the score test. This study propose testing the hypothesis by directly using the maximum likelihood estimators of the mean expression levels in the two groups under negative-binomial distributions. However, with the existence of the dispersion parameter, the calculation becomes difficult and tedious. To ease the computational difficulty, a pseudo likelihood equation and the Stirling's formula approximation are employed. See Piegorsch (1990) and Nakashima (1997). The estimation provides a clear picture of the magnitude of the gene expression. In addition, Nakashima (1997) proved that the resultant estimators have the property of the asymptotic normality. As a consequence, we develop a chi-square test for identifying the differentially expressed genes. Numerical studies, including simulations and real examples, are performed to justify the proposed test.

2 Method

Denote the $G \times 1$ random vector, $Y_{i,j} = (Y_{i,j,1}, \ldots, Y_{i,j,G})^T$, as the G counts of the j-th subject in the i-th phenotypic group, for $j = 1, \ldots, n_i, i = 1, \ldots, k$. Here for simplicity, we consider a binary phenotypic group, $k = 2$. Assume for each $i = 1, 2, g = 1, \ldots, G$, $Y_{i,j,g}$ have the following a negative-binomial distribution,

$$Y_{i,j,g} \sim \ ind. \ NB(\mu_{i,j,g}, \phi_{i,g}), \text{for } j = 1, \ldots, n_i,$$

where $\mu_{i,j,g}$ is the mean parameter and $\phi_{i,g}$ is the over-dispersion parameter. Note that under this distributional model,

$$E(Y_{i,j,g}) = \mu_{i,j,g}, \ \ Var(Y_{i,j,g}) = \mu_{i,j,g} + \phi_{i,g}\mu^2_{i,j,g}.$$

As $\phi_{i,g}$ tends to zero, a negative-binomial distribution becomes a Poisson distribution. Moreover, by taking heterogeneous sequence-depths into account, assume for each $i = 1, 2, j = 1, \ldots, n_i$, there exists $m_{i,j}$ such that

$$\mu_{i,j,g} = m_{i,j}\lambda_{i,g},$$

where $m_{i,j}$ is the depth of the j-th sequence in the group i, and $\lambda_{i,g}$ is the mean relative abundance of the g-th gene in the group i.

The research interest is to determine whether a gene has differential expression levels between the two phenotypic groups. The following hypotheses are tested: For each $g = 1, \ldots, G$,

$$H_{0,g} : \lambda_{1,g} = \lambda_{2,g} \quad \text{v.s.} \quad H_{1,g} : \lambda_{1,g} \neq \lambda_{2,g}.$$

We propose using the Wald's test statistic based on the pseudo maximum likelihood estimators (PMLEs) to test the hypotheses. Let

$$Z_g = \frac{\hat{\lambda}_{1,g} - \hat{\lambda}_{2,g}}{\hat{SE}(\hat{\lambda}_{1,g} - \hat{\lambda}_{2,g})},$$

where $\hat{\lambda}_{i,g}$ are the PMLEs of $\lambda_{i,g}$, and $\hat{SE}(\hat{\lambda}_{1,g} - \hat{\lambda}_{2,g})$ is a consistent estimate of the asymptotic standard error of $\hat{\lambda}_{1,g} - \hat{\lambda}_{2,g}$ for $i = 1, 2, g = 1, \ldots, G$. Asymptotically, under $H_{0,g}$, $Z_g \sim N(0, 1)$. Consequently, given an observed Z-value, z_{g0} the asymptotic p-value is found as $2(1 - \Phi(|z_{g0}|))$, where $\Phi(\cdot)$ is the distribution function of $N(0, 1)$.

Note that in solving the PMLEs, a large observed transcript value of $Y_{i,j,g}$ results in a computational difficulty. From Stirling's formula, the following approximation is employed to overcome the difficulty,

$$\left(\begin{array}{c} N \\ k \end{array} \right) \approx \frac{N^k}{k!}, \ \text{ as } N \to \infty.$$

In a genomic study, the false discovery rate (FDR) is commonly used as an experimentwise error rate for simultaneously testing a numerous number of hypotheses. In literature, the adjusted p-value proposed by Benjamini and Hochberg (1995) and the local false discovery rate proposed by Storey (2003) are two of the most popular multiple testing approaches. In this study, the two approaches are applied on the obtained asymptotic p-values of Z_g's.

3 Results

3.1 Simulations

We consider the simulation study in Auer and Doerge (2011). The transcript levels of ten thousands genes are generated according to the following model:

$$Y_{i,j,g} \overset{\text{indep.}}{\sim} Poisson(\lambda_{i,g}\nu_{i,j,g}), \;\; i = 1, 2, j = 1, \dots, n_i, g = 1, \dots, 10,000,$$

where $\lambda_{i,g}$ is the mean transcript of the $g-$th gene in the group i. Suppose the first eight thousands genes are not differentially expressed in the two groups, while the others are differentially expressed. For each $g = 1, \dots, 10,000$,

$$\lambda_{1,g} \overset{i.i.d.}{\sim} \text{exp(Pareto(location=3, shape=7))}.$$

Furthermore, for $g = 1, \dots, 8000$, $\lambda_{2,g} = \lambda_{1,g}$; for $g = 8001, \dots, 10000$,

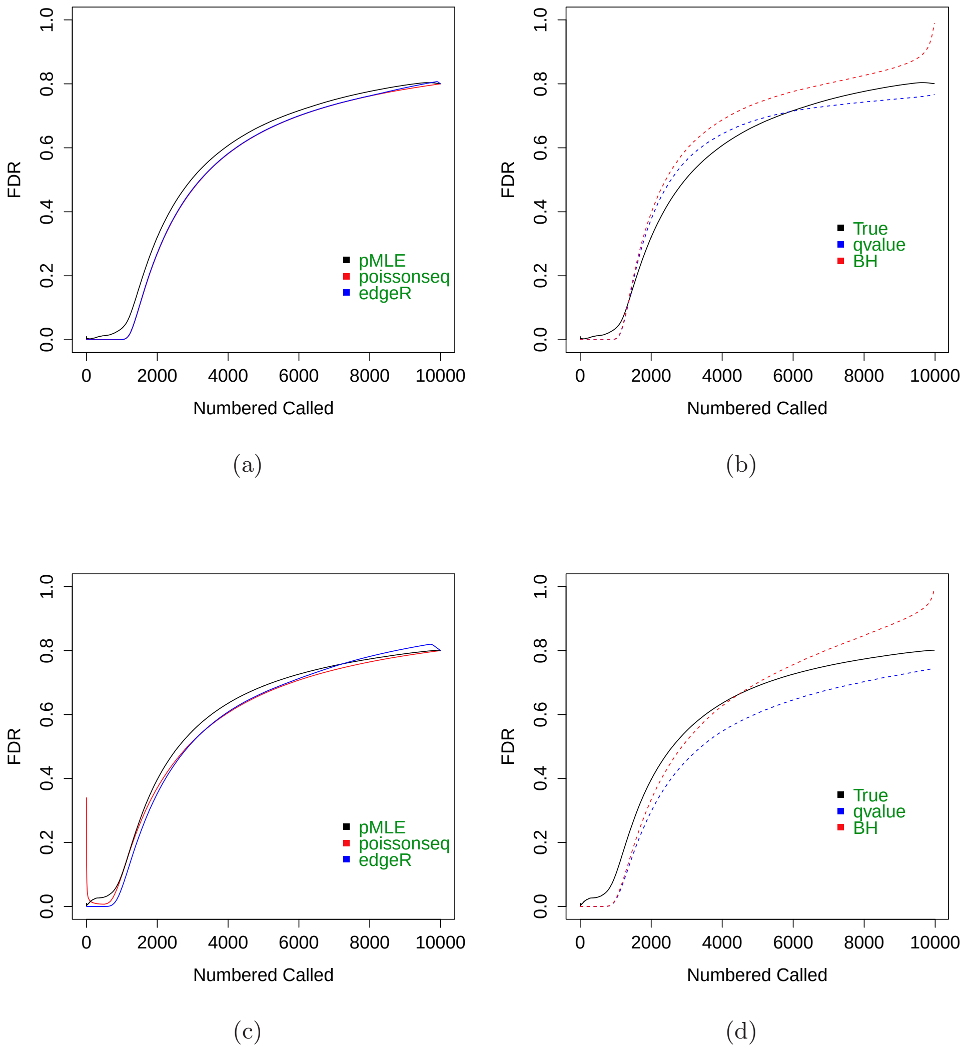
$$\lambda_{2,g} \overset{i.i.d.}{\sim} \text{exp(Pareto(location=3, shape=7))}.$$

On the other hand, $\nu_{i,j,g}$ is the overdispersion parameter. For each gene, we generate a Bernoulli random variable D_g with success probability $p, p \in (0, 1)$ to determine the occurrence of the overdispersion of the gene. The genes of $D_g = 1$ are overdispersed. If $D_{g0} = 1$ for some gene g_0, we then generate $\nu_{i,j,g0}$ from the following Gamma distribution,

$$\nu_{i,j,g0} \overset{i.i.d.}{\sim} Gamma\left(\frac{\lambda_{i,g0}}{\phi - 1}, \frac{\phi - 1}{\lambda_{i,g0}}\right), j = 1, \dots, n_i,$$

for each $i = 1, 2$. If $D_g = 0$, $\nu_{i,j,g} = 1$. We consider $p = 0$ for the scenario without overdispersion, and $p = 0.5$ for the scenario with overdisperion. Consider $n_1 = n_2 = 10$. One hundred replications are generated. The true FDR of the proposed method is estimated via taking an average over the replications. Furthermore, the average over the 100 estimated FDR at each successive rejection is reported as well. The performance of the edgeR method by Robinson et al. (2010) and the PoissonSeq method by Li et al. (2012) are also presented for a comparison. Robinson et al. (2010) suggested the use of the BH adjusted p-value by Benjamini and Hochberg (1995). As well as the BH adjusted p-value, the q-value by Storey (2003) is also employed for our PMLE method.

The results are plotted in Figure 1, in which the top panel considers the scenario without overdispersion and the bottom panel considers the scenario with overdispersion. From (a) and (c), we observe that our test has a slightly higher true FDR than the other two tests. From (b) and (d), both multiplicity adjustments underestimate the true FDR and hence tend to produce liberal results.



(a)

(b)

(c)

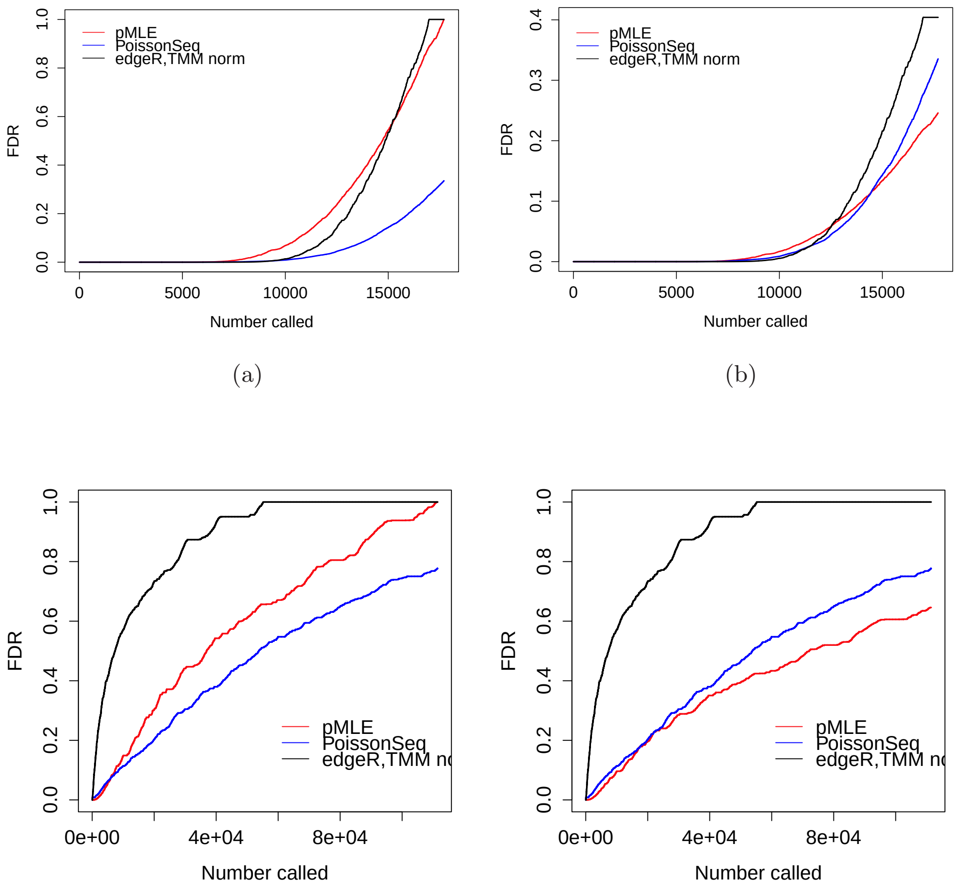
(d)

Figure 1: The top panel gives the FDR curves when the overdispersion does not exist, while the bottom gives the FDR curves when there is an overdispersion. (a), (c): The FDR curves of the three tests. (b), (d) : The average FDR curves of the PMLE test by using the BH adjustment and the q-value (dashed) and the FDR curve (solid).

3.2 Real Examples

The proposed test is applied to the real examples in Marioni et al. (2008) and 't Hoen et al. (2008). Five technical replicates of the liver and of the kidney tissue samples of a single human were sequenced via the Illumina NGS platform. In the example from 't Hoen et al. (2008), four biological replicates of the wild-type and of the transgenic mice were collected. Two estimated FDRs, the BH adjusted p-value by Benjamini and Hochberg (1995) and the q-value by Storey (2003), are employed for the edgeR and our PMLE method. The estimated FDR curves are presented in Figure 2.

The upper panel of Figure 2 are the FDR curves from Marioni et al. (2008), while (a) presents the adjusted p-values by Benjamini and Hochberg (1995), and (b) gives the results of the q-values by Storey (2003). The bottom of Figure 1 are the FDR curves from 't Hoen et al. (2008). In general, the q-value, which adapts an estimate of the proportion of the null hypotheses, is not more conservative than the BH adjusted p-value. We find that using the BH adjustment, the proposed PMLE test and the edgeR provide a conservative result than the PoissonSeq as seen in (a) and (c). However, using the q-value, the PMLE test is more liberal than the edgeR. The PMLE test with the usage of the q-value is comparable with the PoissonSeq in the first example and gives more discoveries in the second example.



(a)

(b)

(c)

(d)

Figure 2: The estimated FDR curves by PMLE, PoissonSeq and edgeR for two data sets: (up) the data set from Marione et al. (2008) and (bottom) the data set from 't Hoen et al. (2008). Moreover, in (a) and (c), the BH-adjusted p-values are used for PMLE and edgeR. In (b) and (d), the q-values are used for PMLE and edgeR.

4 Discussion

In this study, we propose a new test based on the pseudo maximum likelihood estimation to identify the differentially expressed genes in a RNA-Seq data set.

References

[1] Auer, P. L. and Doerge, R. W. (2011) A Two-stage Poisson Model for Testing RNA-Seq Data, *Statistical Applications in Genetics and Molecular Biology*, **10**, Iss. 1, Article 26.

[2] Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing, *J. Roy. Statist. Soc. Ser. B*, **57**, 289-300.

[3] Marioni, J. C., Mason, C.E., Mane, S. M., Stephens, M. and Gilad, Y. (2008) Rna-seq: an Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays, *Genome Res.*, **18**, 1509-1517.

[4] Nakashima, E. (1997) Some Methods for Estimation in a Negative-Binomial Model, *Ann. Inst. Statist. Math.*, **49**, 101-105.

[5] Li, J., Witten, D. M., Johnstone, I. M. and Tibshirani, R. (2012) Normalization, Testing, and False Discovery Rate Estimation for RNA-sequencing Data, *Biostatistics*, **13**, 523-538.

[6] Piegorsch, W. W. (1990) Maximum Likelihood Estimation for the Negative Binomial Dispersion Parameter, *Biometrics*, **46**, 863-867.

[7] Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data, *Bioinformatics*, **26**, 139-140.

[8] Robinson, M. D. and Smyth, G. K. (2007) Moderated Statistical Tests for Assessing Differences in Tag Abundance, *Bioinformatics*, **23**, 2881-2887.

[9] Robinson, M. D. and Smyth, G. K. (2008) Small-sample Estimation of Negative Binomial Dispersion, with Applications to SAGE Data, *Biostatistics*, **9**, 321-332.

[10] Storey, J. D. (2003) The Positive False Discovery Rate: a Bayesian Interpretation and the q-value, *Annals of Statistics*, **31**, 2013-2035.

[11] 't Hoen, P. A. C., Ariyurek, Y., Thygesen, H. H., Vreugdenhil, E., Vossen, R. H., De Menezes, R. X., Boer, J. M., Van Ommen, G. J. and Den Dunnen, J. T. (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research*, **36**, e141.

[12] Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a Revolutionary Tool for Transcriptomics, *Nat. Rev. Genet.*, **10**, 57-63.