A new method for estimating the number of non-differentially expressed genes

J. Wu[1], C.Y. Liu[2], W.T. Chen[2], W.Y. Ma[1] and Y. Ding[1]

[1]Department of Mathematics and Computer Science, Nanjing Medical University, Nanjing, China
[2]Department of Biomedical Engineering, Nanjing Medical University, Nanjing, China

Corresponding author: Y. Ding
E-mail: yongdingcn@126.com

Genet. Mol. Res. 15 (1): gmr.15017402
Received August 8, 2015
Accepted November 26, 2015
Published March 28, 2016
DOI http://dx.doi.org/10.4238/gmr.15017402

ABSTRACT. Control of the false discovery rate is a statistical method that is widely used when identifying differentially expressed genes in high-throughput sequencing assays. It is often calculated using an adaptive linear step-up procedure in which the number of non-differentially expressed genes should be estimated accurately. In this paper, we discuss the estimation of this parameter and point out defects in the original estimation method. We also propose a new estimation method and provide the error estimation. We compared the estimation results from the two methods in a simulation study that produced a mean, standard deviation, range, and root mean square error. The results revealed that there was little difference in the mean between the two methods, but the standard deviation, range, and root mean square error obtained using the new method were much smaller than those produced by the original method, which indicates that the new method is more accurate and robust. Furthermore, we used real microarray data to verify the conclusion. Finally we provide a suggestion when analyzing differentially expressed genes using statistical methods.

Key words: Differentially expressed gene; False discovery rate; Multiple testing

INTRODUCTION

Genomics research has revealed that biological conditions and disease stages are mostly characterized by differences in gene expression levels (DeRisi et al., 1997; Brown and Botstein, 1999; Trapnell et al., 2013). In recent decades, microarray technology has been used as a powerful tool in the quantitative analysis of gene expression (Pollack et al., 2002; Bunney et al., 2003; Smyth, 2004), and the emergence of next-generation sequencing (NGS) technologies has heralded an unprecedented revolution in genome research. RNA sequencing (RNA-Seq), one of the most successful applications of next-generation sequencing technologies, has played an important role in gene expression analysis (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wang et al., 2009). Both microarray data and RNA-Seq data are examples of high-dimensional data characterized by low sample sizes (usually a dozen or dozens) and high dimensionality of variables (genes, typically hundreds, thousands, or even tens of thousands). Therefore, multiple comparisons are needed to identify differentially expressed genes from these data. Traditional methods have controlled the family wise error rate (FWER), which is the probability of committing any type I error. When the number of genes is large, however, the power of detecting differentially expressed genes decreases and truly differentially expressed genes may be missed. In practical research, the main aim is to identify the genes that have significant differences in expression among hundreds of genes; this plays an important role in gene mapping, especially in the recognition of pathogenic genes and the study of disease mechanisms (Burbelo et al., 2014; Rapin et al., 2014). In multiple comparisons, the meaningful index is the expected proportion of incorrectly rejected null hypotheses, as opposed to the probability of even one false discovery. Based on this, the false discovery rate (FDR) approach proposed by Benjamini and Hochberg was a pioneering achievement (Benjamini and Hochberg, 1995). The classical approach requires stringent control of the FWER, with a conservative type I error rate controlled against any configuration of the hypotheses tested. The FDR approach controls the ratio of wrong recognition in an allowable range, and provides a suitable index for multiple testing of large-scale data. Following the seminal paper by Benjamini and Hochberg in 1995, the concept of the FDR has been applied widely in large-scale data analysis. Benjamini and Liu (1999), Benjamini and Yekutieli (2001), Storey (2002), Storey and Tibshirani (2003), Benjamini et al. (2006), and many others have proposed improvements and extensions of the Benjamini and Hochberg method (Kang and Chun, 2011). The adaptive linear step-up (ALSU) procedure proposed by Benjamini et al. in 2006 is the most widely used among all the previous studies. Estimating the number of non-differentially expressed genes is a key step in the ALSU procedure. However, we found that the estimation method proposed in that procedure is not sufficiently accurate; although the mean of the estimates is very close to the true value when the procedure is repeated many times, the standard difference is very large. This introduces large random errors that lead directly to an imprecise final result. In this research, we devised a new method for estimating the number of non-differentially expressed genes, and demonstrated its superiority. The nature of RNA-Seq data has not yet been fully established, and more research is required to understand how these data respond to differential expression analysis (Tarazona et al., 2011). However, microarray technology has proven satisfactory and has wide clinical applications, so we employed microarray data to verify our method. The results could also offer a good reference for analyzing RNA-Seq data.

MATERIAL AND METHODS

Multiple testing and FDR

Multiple testing refers to the simultaneous testing of several hypotheses. A P value test is performed on each hypothesis separately. Consider the problem of testing simultaneously *m* (null) hypotheses, of which m_0 are true and the remaining $m_1 = m - m_0$ are false. Let *V* denote the number of true null hypotheses that are erroneously rejected and let *R* be the total number of hypotheses that are rejected. Table 1 summarizes the situation in a traditional form.

Table 1. Multiple hypotheses testing.

	Not rejected	Rejected	Total
True null	U	V	m_0
Non-true null	T	S	$m - m_0$
	$m - R$	R	m

It is assumed that the specific *m* hypotheses are known in advance. *R* is an observable random variable; *U*, *V*, *S*, and *T* are unobservable random variables. If each individual null hypothesis is tested separately at significance level α, then *R* increases with α. The process of multiple testing is actually the process of multiple comparisons. The primary issue of the process is to set a test criterion. For single hypothesis testing, the criterion usually used is to limit the probability of committing any type I error in a certain range; the test with the smallest probability of making a type II error and the larger test power is then determined. The range of the probability of committing any type I error α is the test level (also called the significance level). Therefore, in a single hypothesis test, the probability of a type I error at level α can be used to control the decision error. However, in multiple hypotheses testing, this approach is invalid. Assuming *m* tests are independent and each individual null hypothesis is tested separately at level α, the probability of making one or more type I error (FWER) will be given by $1-(1-\alpha)^m$. As *m* increases, it tends towards 1. Thus, we must adopt a new approach to control the error in multiple hypotheses testing.

In 1995, Benjamini and Hochberg first proposed the FDR method and the procedure for its control. This new error control theory has attracted much attention from theoretical researchers and application scientists in recent years. FDR is defined as:

$$FDR = \begin{cases} E(V/R), & R > 0 \\ 0, & R = 0 \end{cases}$$

It is the expected proportion of the rejected null hypotheses that have been erroneously rejected. FWER controls errors in the direction of row $m_0 \to V$, while FDR controls errors in the direction of column $R \to V$ in Table 1. This seemingly simple conversion was a significant breakthrough. It not only raises the test power, but also improves traditional multiple hypotheses testing, which is too conservative. Thus, it provides a very appropriate error-measuring criterion for multiple comparisons of large-scale data. Storey and Tibshirani (2003) proposed estimations of m_0 when evaluating the FDR. Benjamini et al. (2006) incorporated this method into the FDR control procedure and the ALSU procedure as follows:

Let H_{0l}, H_{02}, …, H_{0m} be the tested null hypotheses of which H_{0i}, the ith gene, is a non-

differentially expressed gene, and the alternative hypotheses are H_{1i}; the ith gene is a differentially expressed gene. Consider each single test based on the corresponding P values $P_1, P_2, \cdots, P_m$. Let $P_{(1)} \le P_{(2)} \le \cdots \le P_{(m)}$ be the ordered P values, and let the null hypothesis corresponding to $P_{(i)}$ be $H_{0(i)}$. Let each gene belong to either the differentially expressed gene set or the non-differentially expressed gene set, i.e., $H_{0(i)}$ is distributed between 0 and 1. If $P(H_{0(i)} = 1) = \pi_0$, then $P(H_{1(i)} = 1) = 1 - \pi_0$, where π_0 is an unknown parameter of the proportion of true null hypotheses.

Step 1. Let $r(\lambda) = \#\{P_{(i)} \le \lambda\}$, where the number of i that satisfies $P_{(i)} \le \lambda\ (1 \le i \le m)$; λ is usually taken as 0.5.

Step 2. Estimate π_0 by $\hat{\pi}_0 = \frac{m - r(\lambda)}{m(1-\lambda)}$, i.e., the number of true null hypotheses is estimated by $\hat{m}_0 = \frac{m - r(\lambda)}{1-\lambda}$.

Step 3. Estimate $\hat{k} = \max_{1 \le k \le m}\{k : P_{(k)} \le \frac{k}{\hat{m}_0}\gamma\}$ using γ = 0.05 as the critical significance level.

Step 4. If such a $\hat{k}$ exists, reject the $\hat{k}$ hypotheses $H_{0(1)}, H_{0(2)}, \cdots, H_{0(\hat{k})}$; otherwise, do not reject any of the hypotheses.

Step 5. Adjust $P_{(i)}$ as. $\tilde{P}_{(i)} = \min_{i \le k \le m}\{\min(\frac{m}{k}P_{(k)}, 1)\}$

Steps 1 and 2 of the above procedure are used to estimate π_0. Thus, m_0 can be estimated by $\hat{m}_0 = m\hat{\pi}_0$. We found that this approach is very unstable because although the mean of the estimates of m_0 is very close to the true value when the procedure is repeated many times, the standard difference is very large. This causes large random errors because the procedure estimates m_0 only once. As can be seen from the following steps, the estimation of m_0 is a crucial issue in the ALSU procedure, and the accuracy of the estimated value has a significant influence on the identification of differentially expressed genes, the control of the FDR, and the test power. Therefore, it is necessary to improve the estimation method.

Estimation method

If we assume the total number of genes is m, and the non-differentially expressed genes account for π_0, the number of non-differentially expressed genes is $m_0 = m\pi_0$. We can sort the P values, which are obtained by the expression comparison of the two groups of samples of each gene, in ascending order and denote them as $\{P_{(i)}\}$. Thus, the number of genes whose P value is not more than $P_{(i)}$ is simply i. If there are no differentially expressed genes, $P_{(i)}$ is distributed uniformly in [0,1]. Therefore, i and $P_{(i)}$ satisfy the following linear relationship: $i = m_0 P_{(i)}$. Now suppose that there are $m - m_0$ differentially expressed genes. From the statistical point of view, their corresponding P values should be small and less than the significance level α, which leads us to reject the hypothesis H_0. On the other hand, because random errors can lead to type I errors, some non-differentially expressed genes are mistaken for differentially expressed genes. Their P values should also be less than the significance level α. We assume that the P values corresponding to these two classes of genes satisfy $P_{(i)} < \beta$.

Let $n = \max\{i : P_{(i)} < \beta\}$, which is the number of differentially expressed genes we think

of. Then, when $P_{(i)} \geq \beta$, i.e., when $i \geq n+1$, the relationship between i and $P_{(i)}$ is $i = m_0 P_{(i)} + n$. Thus, we take $P_{(i)}$ as the abscissa and i as the ordinate, and fit them with a linear regression. The slope of the regression line is just m_0. Using the least-squares method, we obtained the following equation:

$$m_0 = \frac{\sum_{i=n+1}^{m} iP_{(i)} - \frac{1}{m-n}\sum_{i=n+1}^{m} i \sum_{i=n+1}^{m} P_{(i)}}{\sum_{i=n+1}^{m} P_{(i)}^2 - \frac{1}{m-n}[\sum_{i=n+1}^{m} P_{(i)}]^2} = \frac{\sum_{i=n+1}^{m} iP_{(i)} - \frac{m+n+1}{2}\sum_{i=n+1}^{m} P_{(i)}}{\sum_{i=n+1}^{m} P_{(i)}^2 - \frac{1}{m-n}[\sum_{i=n+1}^{m} P_{(i)}]^2} \quad \text{(Equation 1)}$$

In a practical application, β can be taken as the corresponding probability $P_{(i)}$ when the interval $[\beta,1]$ contains only the probability of non-differentially expressed genes. We can draw the scatter points of i and $P_{(i)}$, and then find the left endpoint of the interval in which i and $P_{(i)}$ show a linear relationship, and take its value as β. In general, β can be taken as slightly larger, because as $P_{(i)}$ increases, the linear relationship between i and $P_{(i)}$ becomes more and more obvious. Although this may lose a few $P_{(i)}$ values, it has little effect on the calculation of m_0 because the number of genes is large. Figure 1 shows the relationship between $P_{(i)}$ and its frequency; $P_{(i)}$ and i are simulated microarray data when $m = 2000$ and π_0 = 0.80 (Table 2). Figure 1a displays the frequency distribution of $P_{(i)}$ in $[0,1]$ with the interval 0.02. Figure 1b is a scatter diagram of i vs. $P_{(i)}$, and shows that β could be taken as 0.05, 0.1, or 0.2. Because of our computation and comparison, we recommend β be taken as twice the significance level α. For example, α is generally taken as 0.05, so β could be taken as 0.1. This is the value we used throughout this article.

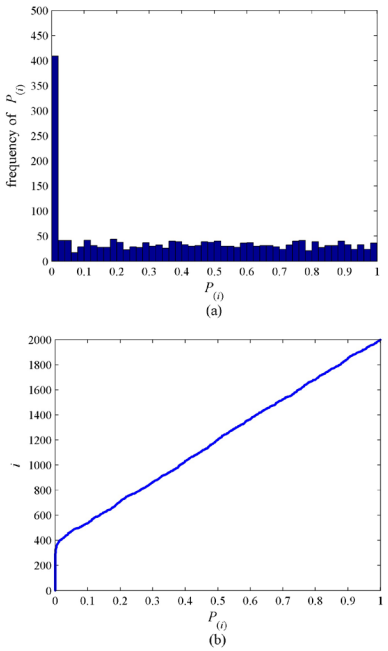

Figure 1. Relationship between $P_{(i)}$ and its frequency; $P_{(i)}$ and i are simulated data (2000 genes). **a.** Frequency distribution of $P_{(i)}$; **b.** $P_{(i)}$ *vs* i.

Table 2. Constitution of the microarray data.

Gene category	Gene number	Sample set S_1	Sample set S_2
		1…40	1…30
Non-differentially expressed genes	Gene 1	$X_{11} \sim N(0,1)$	$X_{12} \sim N(0,1)$
	Gene 2		
	…		
	Gene m_0		
Differentially expressed genes	Gene m_0 +1	$X_{21} \sim N(0,1)$	$X_{22} \sim N(2,1)$
	Gene m_0 +2		
	…		
	Gene m		

Thus, the first two steps of the ALSU procedure can be revised as:

Step 1. Let $n = \max\{i : P_{(i)} < \beta\}$, where β is usually taken as 0.1.

Step 2. When $i \geq n+1$, all the points $(P_{(i)}, i)$ are fitted by linear least squares regression. The slope of the line is used to estimate m_0, i.e., the number of true null hypotheses.

From the ALSU control procedure, we know that the original method only used the number of i that satisfied $P_{(i)} \leq \lambda$. Equation 1 shows that the present method uses more information about $P_{(i)}$, so it should, in theory, provide a better estimate than the original method.

Error analysis

When the number of non-differentially expressed genes is estimated, there will inevitably be some errors that affect the number of differentially expressed genes identified by the ALSU procedure. When the error of estimating m_0 is Δm, what will the change to the number of identified differentially expressed genes k be? We discuss this problem in the following text.

Taking k in the ordered $P_{(k)}$ as the abscissa and the corresponding probability p as the ordinate, then $P_{(k)}$ increases monotonically with k and is noted as $p = P_{(k)}$. The number of differentially expressed genes identified by the ALSU procedure is $\hat{k} = \max_{1 \leq k \leq m}\{k : P_{(k)} \leq \frac{\gamma}{m_0}k\}$ (γ is the FDR level), namely the abscissas of the last intersection point of line $p = \frac{\gamma}{m_0}k$ and curve $p = P_{(k)}$. When the error of estimating m_0 is Δm, the number of identified differentially expressed genes is $\hat{k} + \Delta k = \max_{1 \leq k \leq m}\{k : P_{(k)} \leq \frac{\gamma}{m_0 + \Delta m}k\}$, namely the abscissas of the last intersection point of line $p = \frac{\gamma}{m_0 + \Delta m}k$ and curve $p = P_{(k)}$. When $\Delta m > 0$, we have $\frac{\gamma}{m_0 + \Delta m}k < \frac{\gamma}{m_0}k$, and thus $\hat{k} + \Delta k < \hat{k}$. That is, $\Delta k < 0$ and the number of identified differentially expressed genes decreases. When $\Delta m > 0$, we get the opposite result. Figure 2 shows schematics of the curve $p = P_{(k)}$, the line $p = \frac{\gamma}{m_0}k$, and the line $p = \frac{\gamma}{m_0 + \Delta m}k$ when $\Delta m > 0$.

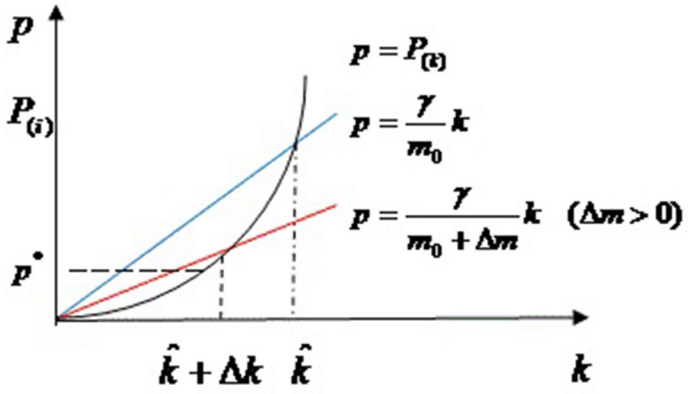

Figure 2. Schematics of curve $p = P_{(k)}$, line $p = \frac{\gamma}{m_0} k$, and line $p = \frac{\gamma}{m_0 + \Delta m} k$.

From Figure 2, we see that at the two intersection points:

$$\Delta p = \left(\frac{k + \Delta k}{m_0 + \Delta m} - \frac{k}{m_0}\right) \gamma \quad \text{(Equation 2)}$$

Thus, the absolute error and relative error of k are:

$$\Delta k = \frac{\Delta p}{\gamma} m_0 + \left(\frac{\Delta p}{\gamma} + \frac{\hat{k}}{m_0}\right) \Delta m \quad \text{(Equation 3)}$$

$$\frac{\Delta k}{\hat{k}} = \frac{\Delta p}{\gamma \hat{k}} (m_0 + \Delta m) + \frac{\Delta m}{m_0} \quad \text{(Equation 4)}$$

If the number of identified differentially expressed genes is $\hat{k} = m - m_0$, and the corresponding ordinate is p^*, we get $p^* = \frac{\gamma}{m_0}(m - m_0)$. Thus,

$$\gamma = \frac{m_0 p^*}{m - m_0} \quad \text{(Equation 5)}$$

This is simply the FDR level. From the definition of FDR and the equation $\frac{V}{R} = \gamma$, we know that when $R = m - m_0$, the number of differentially expressed genes that are wrongly identified is:

$$V = m_0 p^* \quad \text{(Equation 6)}$$

In general, there are fewer differentially expressed genes than non-differentially expressed genes. We can see from Table 1 that when $R = m - m_0$, the FDR level $FDR = \frac{V}{m - m_0}$ is larger than the probability of making a type I error $\alpha = V / m_0$. The relationship between them is $\frac{FDR}{\alpha} = \frac{m_0}{m - m_0} = \frac{\pi_0}{1 - \pi_0}$. For

example, when $\pi_0 = 0.8$, the ratio is four. Therefore, we can control the probability of committing a type I error by controlling the FDR level. We can also have:

$$\alpha = \frac{m - m_0}{m_0} FDR = \frac{1 - \pi_0}{\pi_0} FDR \quad \text{(Equation 7)}$$

That is, we can compute the probability of committing a type I error using the FDR level. For example, when the *FDR* = 0.05 and $\pi_0 = 0.8$, we get $\alpha = 0.0125$.

On the other hand, when $R = m - m_0$ and $T = V$, the probability of committing a type II error is:

$$\beta = \frac{T}{m - m_0} = \frac{V}{m - m_0} = FDR \quad \text{(Equation 8)}$$

The power of the statistical test $\widetilde{T}$, which is the probability that the test correctly rejects the null hypothesis (H_0) when the alternative hypothesis (H_1) is true, is:

$$\widetilde{T} = \frac{S}{m - m_0} = \frac{R - V}{m - m_0} \quad \text{(Equation 9)}$$

when $R = m - m_0$, we get:

$$\widetilde{T} = \frac{m - m_0 - V}{m - m_0} = 1 - \frac{m_0 \alpha}{m - m_0} = 1 - \frac{\pi_0}{1 - \pi_0} \alpha \quad \text{(Equation 10)}$$

RESULTS

Since microarray technology has proven adequate for clinical applications, we employed microarray data to prove the method described above. The results could also provide a good reference for analyzing RNA-Seq data.

Performance on simulated data

We analyzed simulated microarray data to estimate m_0 using the original and new methods, and tested and compared the results. First, we simulated the data from a gene chip. For convenience, we denoted the sample set of healthy people as S_1 (comprising 40 samples) and the sample set of cancer patients as S_2 (comprising 30 samples); the total number of genes was 2000 ($m = 2000$). For a non-differentially expressed gene, the microarray data in S_1 and S_2 should have the same distribution, which is usually supposed to be a normal distribution: $N(0,1)$. On the other hand, for a differentially expressed gene, the microarray data in S_1 and S_2 should have different distributions. We also assumed $N(0,1)$ in S_1 but $N(2,1)$ in S_2. The number of non-differentially expressed genes could be $m_0 = m\pi_0$. We drew m_0 samples with a $N(0,1)$ distribution and a capacity of 40, which was denoted as X_{11}. We also collected m_0 samples with

an $N(0,1)$ distribution and a capacity of 30 which was denoted as X_{12}. X_{11} and X_{12} constituted the non-differentially expressed gene data. We then drew $m-m_0$ samples with an $N(0,1)$ distribution and a capacity of 40, and denoted them as X_{21}. We also generated $m-m_0$ samples with an $N(2,1)$ distribution and a capacity of 30, and denoted them as X_{22} . X_{21} and X_{22} constituted the differentially expressed gene data. Thus, we derived a matrix with 2000 rows and 70 columns and utilized it as the simulated data of a gene chip (Table 2).

For each gene, the expression data in sets S_1 and S_2 were compared by *t*-test and the corresponding P values were sorted in ascending order. We then estimated m_0 with the ALSU procedure using the original and new methods. To ensure that the comparison results had statistical significance, we repeated the simulation 1000 times for different values of π_0 and analyzed the value of 1000 $\hat{m}_0$ calculated using the two methods.

To evaluate the accuracy of the estimated value, we used the following four indicators: mean, standard deviation (SD), range, and root mean square error (RMSE). The mean reflects the concentration trend of the data. The SD represents the discrete degree of the data. The range is the difference between the maximum and minimum values in a list of numbers, and can reflect the fluctuation range and the discrete degree of the data. The RMSE is defined as $\sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}}$ and is used to measure the differences between the estimated values and the true values, in which d_i is the deviation between the estimated value and the true value for the ith estimate. This index is very sensitive to large errors; therefore, it is a good measure of accuracy. The smaller the index, the more accurate the estimated values are.

We took the value of π_0 on the interval [0.7, 0.95] with the step 0.05 and computed the mean, SD, range, and RMSE of each m_0 estimate corresponding to each π_0. Table 3 summarizes the estimates.

Table 3 shows that the mean of m_0 estimated using these two methods is very close to the true value. However, the SD, range, and RMSE derived using the new method are about 60% of the values obtained using the original method. This shows that the discrete extent of the data calculated using the proposed method is smaller, and the estimate is closer to the true value. Thus, the algorithm is more stable and robust.

Table 3. Comparison of results of m_0 estimation using the two methods when $X_{22} \sim N(2,1)$.

π_0	m_0	Mean (original method/ proposed method)	SD (original method/ proposed method)	Range (original method/ proposed method)	RMSE (original method/ proposed method)
0.70	1400	1399.1/1399.7	38.37/21.63	268/140.5	38.36/21.62
0.75	1500	1500.7/1500.5	38.95/22.35	240/142.5	38.93/22.35
0.80	1600	1600.2/1600.3	41.11/22.98	274/147.6	41.09/22.97
0.85	1700	1698.4/1699.6	41.19/23.87	254/153.3	41.20/23.86
0.90	1800	1799.3/1800.2	41.61/24.80	272/161.6	43.59/24.79
0.95	1900	1902.2/1901.0	44.37/24.80	294/170.2	44.39/24.81

SD = standard deviation; RMSE = root mean square error.

We chose $\pi_0 = 0.80$ for further analysis and obtained the estimates of m_0 with the two methods by analyzing 1000 simulations. Testing using the function “lillietest” in the MATLAB software indicated that the two groups of data conformed to normal distribution. Figure 3 shows the frequency distribution diagram and box plot of $\hat{m}_0$ on the interval [1480, 1720] with a distance of 10.

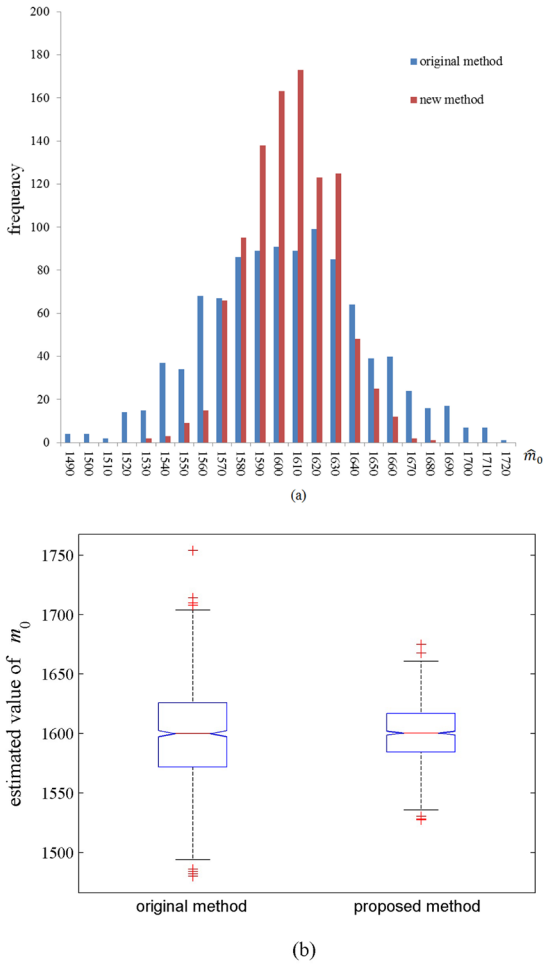

Figure 3. Values of $\hat{m}_0$ for the new and original methods when $\pi_0 = 0.80$. **a.** Frequency distribution diagram; **b.** box plots comparing the original and proposed methods.

From Table 3 and Figure 3, we observed that the 1000 $\hat{m}_0$ estimated using the new method conforms better to a normal distribution, with a mean value of 1600 and a smaller SD. Figure 3 also shows that the value of m_0 estimated by the two methods is symmetrical; however, the estimations calculated using the new method are more concentrated and the number of abnormal values is significantly lower than that found using the original method. These results show that the estimates calculated using the new method were generally concentrated near the true value, and their SD, range, and RMSE were greatly reduced. Therefore, the results calculated using the new method are more accurate and the new method is more stable.

In the simulation mentioned above, we assumed the distributions of the differentially expressed and the non-differentially expressed genes were $N(0,1)$ and $N(2,1)$ respectively. In general, if two normal distributions have the same variance, the closer means would make it more difficult to distinguish between them. To further test the validity of the new method, we also carried

out the same simulation experiments when X_{22} was subjected to distribution $N(1,1)$ (the others parameters were not changed). The results are shown in Table 4. The results in Table 4 are similar to those in Table 3, which further illustrates that the new method provides better discrimination.

Table 4. Comparison of results of m_0 estimation using the two methods when $X_{22} \sim N(1,1)$.

π_0	m_0	Mean (original method/ proposed method)	SD (original method/ proposed method)	Range (original method/ proposed method)	RMSE (original method/ proposed method)
0.70	1400	1401.5/1403.3	36.93/21.42	218/132.4	36.95/21.66
0.75	1500	1500.4/1502.9	38.25/22.71	266/131.0	38.24/22.88
0.80	1600	1599.0/1600.8	39.98/24.16	232/144.6	39.97/24.16
0.85	1700	1701.6/1700.8	40.67/23.78	226/149.4	40.68/23.78
0.90	1800	1797.9/1800.9	42.21/24.50	268/154.4	42.24/24.50
0.95	1900	1902.7/1899.5	44.71/25.20	296/186.6	44.77/25.19

SD = standard deviation; RMSE = root mean square error.

Performance on real data

We also employed prostate cancer gene expression profiles to verify the new method. The dataset was downloaded from the National Center for Biotechnology Information Gene Expression Omnibus database (Barrett et al., 2005) (accession number: GSE6919). We selected 83 samples from the database, which included 18 normal samples and 65 primary tumor samples; each sample had 12,625 genes. The raw data were preprocessed using a robust multi-array average (RMA) procedure (Irizarry et al., 2003). The expression levels of the two groups of each gene were tested for homogeneity of variance and approximately 73% of genes passed the test. We computed the P value for each gene. If a gene passed the test, we used a t-test. Otherwise, we used Satterthwaite's approximate t-test. We sorted the P values into ascending order. Setting the FDR control level to 0.05, we estimated m_0, which represents the number of non-differentially expressed genes, using the original and new methods, and identified the differentially expressed genes associated with prostate cancer using the FDR method. The results are shown in Table 5.

Table 5. Calculated results of microarray data from prostate cancer samples.

Dataset	Accession number	Number of samples*	Number of genes	$\hat{m}_0$		No. of differentially expressed genes	
				Original method	Proposed method	Original method	Proposed method
Chandran [21]	GSE6919	18:65	12,625	7760	8845	2133	1968

*normal *vs* tumor.

Table 5 indicates that, compared with the original method, the number of m_0 estimated by the new method increased by 1085 (14%), which reduced the number of differentially expressed genes by 165 (7.7%). This shows that the estimation of m_0 influences the identification of differentially expressed genes using the FDR method. Not only is our method more accurate in theory, it also has practical value in the calculation of real microarray data.

From the value of $P_{(i)}$, we know $P_{(1968)} = 0.0111$, $P_{(2133)} = 0.0137$, and $\Delta p = -0.0026$. Taking $m_0 = 7760$, $\Delta m = 1085$, $k = 2133$, and $\gamma = 0.05$, we obtained $\Delta k = -162$ using Equation 2. This result is very close to the reduction in the number of differentially expressed genes given in Table 5 (165), which further verified the accuracy of Equation 2.

The number of non-differentially expressed genes m_0 estimated by the proposed method

was 8845, so the number of differentially expressed genes was 12,625 - 8845 = 3780. If the number of identified differentially expressed genes is 3780, the corresponding P value is $P_{(3780)} = 0.0675$. Using Equations 5-10, we can determine that the FDR level was 0.158. So the number of wrongly identified genes was 597 out of a total number of 3780, and the probability of committing any type I errors α was approximately 0.07. The test power was $T = 0.842$ and the probability of committing any type II errors was 0.158.

DISCUSSION

Estimating the number of true null hypotheses is one of the crucial issues in multiplicity testing (Kang and Lee, 2012). We have improved the stability of the estimation method in the classical ALSU procedure and present the new method here. The estimation results of the two methods were compared using simulated microarray data with mean, SD, range, and RMSE as evaluation indices. The new method was also verified using data from actual gene expression profiles. All the results indicated that our method effectively improved the stability and accuracy of the estimation of m_0. Thus, it further improved the accuracy of identifying differentially expressed genes using the FDR method.

The existing methods to identify differentially expressed genes determine the FDR level γ first, and then identify differentially expressed genes using the FDR method according to γ (Rapaport et al., 2013; Colquhoun, 2014; Frazee et al., 2014). It is important to note that the number of differentially expressed genes identified by this method will vary with different values of γ, which is not the true number of differentially expressed genes. For example, for the above prostate cancer microarray data, we identified 1968 differentially expressed genes when $\gamma = 0.05$. While selecting $\gamma = 0.01$, we identified 908 differentially expressed genes. However, in fact the data contain about 3780 differentially expressed genes by our estimation. Therefore, we suggest that we should add the following content when analyzing differentially expressed genes using statistical methods: if we have a method that can estimate the number of differentially expressed genes accurately, we can use this number to calculate the identification criteria. The FDR method can then conversely be used to determine the FDR level, the number of misidentifications, the probability of committing any type I or type II errors, and the power of the statistical test. All of these issues can be solved using Equations 5-10. Thus, the number of differentially expressed genes identified by the new method is consistent with the true situation. We also obtained a general understanding of the number of misidentifications and other information. This provides us with a more comprehensive understanding of identification. Moreover, when the FDR level is controlled, the probability of committing type I errors is also controlled. Traditional methods control the errors in the direction of row $m_0 \to V$, and the FDR controls the error in the direction of column $R \to V$ in Table 1. If this seemingly simple conversion is a breakthrough in statistics, it follows that the above suggestion, which first determines the number of differentially expressed genes and then determines the test level, is also a breakthrough for determining the number of differentially expressed genes.

Conflict of interests

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors thank Zhengrui Xiao for suggestions and corrections that improved the text.

Research supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (#13KJB310007), and the Science and Technology Development Fund Key Project of Nanjing Medical University (#2013NJMU006).

REFERENCES

Barrett T, Suzek TO, Troup DB, Wilhite SE, et al. (2005). NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res.* 33: D562-D566. http://dx.doi.org/10.1093/nar/gki022

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.R. Stat. Soc.* 57: 289-300.

Benjamini Y, Krieger AM and Yekutieli D (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 93: 491-507. http://dx.doi.org/10.1093/biomet/93.3.491

Benjamini Y and Liu W (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Stat. Plan. Inference* 82: 163-170. http://dx.doi.org/10.1016/S0378-3758(99)00040-3

Benjamini Y and Yekutieli D (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29: 1165-1188.

Brown PO and Botstein D (1999). Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21 (Suppl): 33-37. http://dx.doi.org/10.1038/4462

Bunney WE, Bunney BG, Vawter MP, Tomita H, et al. (2003). Microarray technology: a review of new strategies to discover candidate vulnerability genes in psychiatric disorders. *Am. J. Psychiatry* 160: 657-666. http://dx.doi.org/10.1176/appi.ajp.160.4.657

Burbelo PD, Ambatipudi K and Alevizos I (2014). Genome-wide association studies in Sjögren's syndrome: What do the genes tell us about disease pathogenesis? *Autoimmun. Rev.* 13: 756-761. http://dx.doi.org/10.1016/j.autrev.2014.02.002

Colquhoun D (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *R. Soc. Open Sci.* DOI: 10.1098/rsos.140216.

DeRisi JL, Iyer VR and Brown PO (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278: 680-686. http://dx.doi.org/10.1126/science.278.5338.680

Frazee AC, Sabunciyan S, Hansen KD, Irizarry RA, et al. (2014). Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics* 15: 413-426. http://dx.doi.org/10.1093/biostatistics/kxt053

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4: 249-264. http://dx.doi.org/10.1093/biostatistics/4.2.249

Kang M and Chun H (2011). A generalized false discovery rate in microarray studies. *Comput. Stat. Data Anal.* 55: 731-737. http://dx.doi.org/10.1016/j.csda.2010.06.017

Kang M and Lee J (2012). Bayesian inference for the proportion of true null hypotheses using minimum Hellinger distance. *J. Stat. Plan. Inference* 142: 820-825. http://dx.doi.org/10.1016/j.jspi.2011.10.001

Mortazavi A, Williams BA, McCue K, Schaeffer L, et al. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5: 621-628. http://dx.doi.org/10.1038/nmeth.1226

Nagalakshmi U, Wang Z, Waern K, Shou C, et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344-1349. http://dx.doi.org/10.1126/science.1158441

Pollack JR, Sørlie T, Perou CM, Rees CA, et al. (2002). Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA* 99: 12963-12968. http://dx.doi.org/10.1073/pnas.162471999

Rapaport F, Khanin R, Liang Y, Pirun M, et al. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* 14: R95. http://dx.doi.org/10.1186/gb-2013-14-9-r95

Rapin N, Bagger FO, Jendholm J, Mora-Jensen H, et al. (2014). Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients. *Blood* 123: 894-904. http://dx.doi.org/10.1182/blood-2013-02-485771

Smyth GK (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3: e3. http://dx.doi.org/10.2202/1544-6115.1027

Storey JD (2002). A direct approach to false discovery rates. *J.R. Stat. Soc.* 64: 479-498. http://dx.doi.org/10.1111/1467-9868.00346

Storey JD and Tibshirani R (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* 100: 9440-9445. http://dx.doi.org/10.1073/pnas.1530509100

Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, et al. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Res.* 21: 2213-2223. http://dx.doi.org/10.1101/gr.124321.111

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, et al. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* 31: 46-53. http://dx.doi.org/10.1038/nbt.2450

Wang Z, Gerstein M and Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57-63. http://dx.doi.org/10.1038/nrg2484