View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Crossref

Li *et al. BMC Bioinformatics* (2017) 18:1
DOI 10.1186/s12859-016-1414-x

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Bon-EV: an improved multiple testing procedure for controlling false discovery rates

Dongmei Li[1]*, Zidian Xie[2], Martin Zand[1], Thomas Fogg[1] and Timothy Dye[1,3]

Abstract

Background: Stability of multiple testing procedures, defined as the standard deviation of total number of discoveries, can be used as an indicator of variability of multiple testing procedures. Improving stability of multiple testing procedures can help to increase the consistency of findings from replicated experiments. Benjamini-Hochberg's and Storey's *q*-value procedures are two commonly used multiple testing procedures for controlling false discoveries in genomic studies. Storey's *q*-value procedure has higher power and lower stability than Benjamini-Hochberg's procedure. To improve upon the stability of Storey's *q*-value procedure and maintain its high power in genomic data analysis, we propose a new multiple testing procedure, named Bon-EV, to control false discovery rate (FDR) based on Bonferroni's approach.

Results: Simulation studies show that our proposed Bon-EV procedure can maintain the high power of the Storey's *q*-value procedure and also result in better FDR control and higher stability than Storey's *q*-value procedure for samples of large size(30 in each group) and medium size (15 in each group) for either independent, somewhat correlated, or highly correlated test statistics. When sample size is small (5 in each group), our proposed Bon-EV procedure has performance between the Benjamini-Hochberg procedure and the Storey's *q*-value procedure. Examples using RNA-Seq data show that the Bon-EV procedure has higher stability than the Storey's *q*-value procedure while maintaining equivalent power, and higher power than the Benjamini-Hochberg's procedure.

Conclusions: For medium or large sample sizes, the Bon-EV procedure has improved FDR control and stability compared with the Storey's *q*-value procedure and improved power compared with the Benjamini-Hochberg procedure. The Bon-EV multiple testing procedure is available as the BonEV package in R for download at https://CRAN.R-project.org/package=BonEV.

Keywords: Multiple testing procedure, FDR, Power, Stability, RNA-Seq

Background

Microarray and next-generation sequencing (NGS) technologies have been widely used in biological and biomedical research to identify novel biomarkers and genomic modifications related to biological processes and diseases. Multiple testing procedures are widely used in microarray and NGS studies to control the multiple testing error rate to minimize false discoveries from the enormous number of simultaneous hypothesis tests [1]. Many multiple testing error rates have been proposed such as family-wise error rate (FWER) [2], *k* family-wise error rate (kFWER) [3], and false discovery rate (FDR) [4]. For discovery purposes, the false discovery rate (FDR), defined as the expected proportion of false discoveries among total number of discoveries, is often controlled in multiple testing procedures to select significant features in microarray and NGS studies [4–6]. Benjamini and Hochberg's FDR controlling procedure [4] and Storey's *q*-value procedure [7, 8] are the most commonly used procedures [9]. The Bonferroni procedure, although perceived as a conservative procedure for multiple testing error rate control, has stability superior to Benjamini-Hochberg's procedure in terms of variability of total number of discoveries, and equivalent power to Benjamini-Hochberg's procedure, when used to control the expected number of false discoveries (EV,

*Correspondence: dongmei_li@urmc.rochester.edu
[1] Clinical and Translational Science Institute, School of Medicine and Dentistry, University of Rochester, 265 Crittenden Boulevard CU 420708, 14642 Rochester, NY, USA
Full list of author information is available at the end of the article

BioMed Central

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

where V stands for the number of false discoveries) at a user-specified level [10].

In this study, we examine the stability (defined as standard deviation of the total number of rejected hypotheses) of both Benjamini-Hochberg's FDR controlling procedure and Storey's q-value procedure for generating adjusted p-values to select significant genes or biomarkers in microarray and NGS data analysis. In addition, we propose our own multiple testing procedure (named Bon-EV) based on Bonferroni's EV controlling procedure, that has equivalent power, higher stability, and better FDR control than the Storey's q-value procedure with at least medium-sized samples in microarray and NGS data analysis. Multiple testing procedures with high power, good FDR control, and high stability are desirable in genomic data analysis due to the high cost of sequencing in genomic studies. The Bon-EV multiple testing procedure will be attractive to genomic data analysts as it not only maintains the high power of Storey's q-value procedure, but also offers better FDR control and higher stability, especially for small to medium sample size studies that need high stability, high power and good FDR control to maximize the odds of true discoveries.

Methods

Suppose we are testing m null hypotheses simultaneously in a high-dimensional data analysis for single nucleotide polymorphism (SNP) identification, differential gene expression, or DNA methylation discovery. Among m null hypotheses, m_0 null hypotheses are true null hypotheses. Among R rejected null hypotheses, V hypotheses are false rejections (false discoveries). Multiple testing error rates need to be controlled to minimize false discoveries among total rejections. False discovery rate (FDR) is a commonly used multiple testing error rate in genomic analysis. Several definitions of FDR have been proposed to measure the false discovery rate such as FDR, positive false discovery rate (pFDR), and $\frac{E(V)}{E(R)}$. The FDR and pFDR are defined as:

$$FDR = E\left(\frac{V}{R}|R > 0\right) Pr(R > 0), \tag{1}$$

$$pFDR = E\left(\frac{V}{R}|R > 0\right). \tag{2}$$

Benjamini-Hochberg procedure

The Benjamini-Hochberg procedure [4] provides control of FDR at level α through the following step-up procedure:

- Order original p-values $p_i, i = 1, \ldots, m$, from the smallest to the largest such that $p_{(1)} \leq p_{(2)} \cdot \leq p_{(m)}$;
- Find k as the largest i for which $P_{(i)} \leq \frac{i}{m}\alpha$;
- Reject all null hypotheses $H_i, i = 1, 2, \ldots, k$.

The Benjamini-Hochberg procedure provides $FDR = \frac{m_0}{m}\alpha \leq \alpha$, a strong control for FDR at level α for independent and positively correlated test statistics. Meanwhile, the Benjamini-Hochberg procedure is also conservative by a factor of $\frac{m_0}{m}$ for controlling FDR at level α.

Storey's *q*-value procedure

Arguing that where $m_0 = m$, one would not be interested in cases where no test is significant ($FDR = 1$ in this situation), Storey [7] proposes the definition of positive false discovery rate (pFDR) that is conditional on at least one rejection. The Storey's q-value procedure used for controlling pFDR improves power over the Benjamini-Hochberg procedure by including the estimation of $\pi_0 = \frac{m_0}{m}$. Storey's q-value procedure proceeds as follows:

- Order original p-values $p_i, i = 1, \ldots, m$, from the smallest to the largest such that $p_{(1)} \leq p_{(2)} \cdot \leq p_{(m)}$;
- Find k as the largest i for which $P_{(i)} \leq \frac{i}{m\hat{\pi_0}}\alpha$ where $\pi_0 = \frac{m_0}{m}$;
- Reject all null hypothesis $H_i, i = 1, 2, \ldots, k$.

Storey proposes to estimate π_0 conservatively by $\hat{\pi_0} = \frac{\sharp\{p_i > \lambda\}}{(1-\lambda)m}$, where λ is chosen to minimize the mean-squared error of the $pFDR$ estimates.

Bonferroni procedure

The Bonferroni procedure has traditionally been considered too conservative for genomic data analysis for discovery purposes. Gordon et al. [10] show the Bonferroni procedure has comparable power and superior stability to the Benjamini-Hochberg procedure when used to control the expected number of false discoveries ($E(V)$). The Bonferroni procedure rejects H_i if $p_i \leq \frac{\gamma}{m}$, and controls $E(V)$ at a pre-specified number of tests γ when test statistics are either independent or correlated. To prove that the Bonferroni procedure controls $E(V)$ at level γ, we assume p_i for $i = 1, 2, \ldots, m_0$ has an independent uniform distribution, then we have

$$E(V) = \sum_{i=1}^{m_0} Pr\left(p_i \leq \frac{\gamma}{m}\right) = m_0\frac{\gamma}{m} \leq \gamma. \tag{3}$$

If we let $\gamma = \alpha \cdot E(R)$, then we have $E(V) \leq \alpha \cdot E(R)$ and $\frac{E(V)}{E(R)} \leq \alpha$. Notice that the Bonferroni procedure used to control $E(V)$ is conservative by a factor of $\frac{m_0}{m}$. Thus, we further improve power of the Bonferroni procedure by estimating m_0 and replacing m with a conservative estimator of m_0 in the cutoff value.

Bon-EV procedure

Based on theorem 1 in Storey's 2003 paper, we propose our own Bon-EV procedure to control the $pFDR$ at level α. Theorem 1 states that: Suppose that m identical hypothesis tests are performed with p-values $P_1, \ldots, P_m$ and significant region includes all adjusted p-values P^* less than or equal to α. Assume that (P_i, H_i) are *i.i.d.* random variables, $P_i|H_i \sim (1-H_i)\cdot F_0 + H_i \cdot F_1$ for some null distribution F_0 and alternative distribution F_1, and $H_i \sim Bernoulli(\pi_1)$ for $i = 1, \ldots, m$. Then

$$pFDR = Pr(H=0|P^* \le \alpha) \tag{4}$$

Based on $pFDR = Pr(H = 0|P^* \le \alpha)$, our Bon-EV procedure is as follows:

- Compare each p_i with $\frac{\alpha\cdot Pr(\widehat{P^*}\le\alpha)}{\hat{\pi}_0}$, $i = 1, 2, \ldots, m$;
- Reject H_i if $p_i \le \frac{\alpha\cdot Pr(\widehat{P^*}\le\alpha)}{\hat{\pi}_0}$, $i = 1, 2, \ldots, m$.

In our Bon-EV procedure, P^* are the adjusted p-values from the Benjamini-Hochberg procedure, and $Pr(P^* \le \alpha)$ is estimated by the proportion of null hypotheses that have the Benjamini-Hochberg adjusted p-values $\le \alpha$. We use the same estimate of π_0 as used in Storey's q-value procedure.

The following equations show that our Bon-EV procedure controls $pFDR$ at level α.

$$V = \sum_{i=1}^{m} I(P^* \le \alpha \; and \, H_i = 0). \tag{5}$$

$$\begin{aligned} E(V) &= \sum\nolimits_{i=1}^{m} Pr(P_i^* \le \alpha \; and \quad H_i = 0) \\ &= \sum\nolimits_{i=1}^{m} Pr(H_i = 0|P_i^* \le \alpha) Pr(P_i^* \le \alpha) \\ &= mPr(H=0|P^* \le \alpha) Pr(P^* \le \alpha) \\ &= pFDR \cdot (mPr(P^* \le \alpha)). \end{aligned} \tag{6}$$

So, if $E(V) \le \alpha \cdot (mPr(P^* \le \alpha))$, then $pFDR \le \alpha$. To have $E(V) \le \alpha \cdot (mPr(P^* \le \alpha))$ using the Bonferroni approach, we compare each p_i with $\frac{\alpha\cdot(mPr(P^*\le\alpha))}{m_0}$:

$$\begin{aligned} E(V) &= \sum\nolimits_{i=1}^{m_0} Pr\left(p_i \le \frac{\alpha\cdot(mPr(P^*\le\alpha))}{m_0}\right) \\ &= \sum\nolimits_{i=1}^{m_0} Pr\left(p_i \le \frac{\alpha\cdot(mPr(P^*\le\alpha))}{m\cdot\pi_0}\right) \\ &= m_0 \cdot \frac{\alpha\cdot(mPr(P^*\le\alpha))}{m_0} \\ &= \alpha \cdot (mPr(P^* \le \alpha)). \end{aligned} \tag{7}$$

Thus, $pFDR$ will be controlled at level α if each p_i is compared with $\frac{\alpha\cdot(mPr(P^*\le\alpha))}{m_0}$. As $m_0 = m\cdot\pi_0$, we compare each p_i with $\frac{\alpha\cdot Pr(P^*\le\alpha)}{\pi_0}$. $Pr(P^* \le \alpha)$ is estimated by $Pr(\widehat{P^*} \le \alpha) = \frac{\sharp\{P_i^*\le\alpha\}}{m}$ from Benjamini-Hochberg's FDR controlling method and π_0 is estimated by $\hat{\pi}_0 = \frac{\sharp\{p_i>\lambda\}}{(1-\lambda)m}$ from Storey's q-value method. The expected values of $Pr(\widehat{P^*} \le \alpha)$ and $\hat{\pi}_0$ are

$$\begin{aligned} E(Pr(\widehat{P^*} \le \alpha)) &= E\left[\frac{\sharp\{P_i^*\le\alpha\}}{m}\right] \\ &= Pr(P^* \le \alpha); \end{aligned} \tag{8}$$

$$E(\hat{\pi}_0) = E\left[\frac{\sharp\{p_i > \lambda\}}{(1-\lambda)m}\right] = \frac{(1-\lambda)m}{(1-\lambda)m} = 1 \ge \pi_0. \tag{9}$$

Thus, our procedure controls $E(V)$ at $\alpha \cdot (mPr(P^* \le \alpha))$ and controls $pFDR$ at level α. We took advantage of existing R functions to estimate $Pr(P^* \le \alpha)$ and π_0. We estimate $Pr(P^* \le \alpha)$ from the *p.adjust* function in R with Benjamini and Hochberg's method and estimate π_0 using the *qvalue* and *pi0est* function from the *qvalue* package.

Our proposed Bon-EV procedure integrates the approaches of the Bonferroni procedure, Benjamini-Hochberg procedure, and Storey's q-value procedure. The estimated π_0 from Storey's q-value procedure used in the Bon-EV procedure increases the cutoff value for p-values in each comparison, thus improving its power compared to Benjamini-Hochberg's procedure. As the sample size increases, the $\hat{\pi}_0$ is closer to the value of π_0, and the power will be further improved. By adapting the estimate of $Pr(P^* \le \alpha)$ from the Benjamini-Hochberg procedure for the Bon-EV procedure, the proportion of false discoveries is reduced compared to Storey's q-value procedure. Thus, we expect the Bon-EV procedure to have better FDR control than Storey's q-value procedure. Regarding the stability, single-step approaches such as the Bonferroni and Bon-EV procedures are superior to stepwise approaches such as the Benjamini-Hochberg procedure and Storey's q-value procedure, but, the inclusion of the π_0 estimate in the Bon-EV and Storey's q-value procedures reduces the stability. Taken together, we expect the Bon-EV procedure to have better stability than Storey's q-value procedure.

Results

Simulation studies

We conduct simulation studies to evaluate the FDR control, power, and stability of our Bon-EV procedure, the Benjamini-Hochberg procedure and Storey's q-value procedure. Power is defined as the proportion of true rejections among total non-true null hypotheses, and stability is defined as the standard deviation (SD) of total number of rejections [11].

$$Power = E\left(\frac{S}{m - m_0}\right), \tag{10}$$

$$Stability = SD(R). \tag{11}$$

Our simulations compare gene expression levels between two groups with equal sample sizes of 5, 15, and 30 in each group. For each sample, we test 10,000 genes with expression levels following multivariate normal distributions with means at vector of 0 for the control group and means at vector of $(\mu, 0)$ for the treatment group. We set standard deviations at 1 with correlations between genes equal to ρ (ρ = 0, 0.4, or 0.8, depending on the simulation study). All genes were equally correlated with correlation ρ. Meanwhile, we also conduct simulations with pairwise gene correlations randomly selected from uniform(0, 0.8) distribution for sample sizes of 5, 15, and 30 in each of the two groups. We set the proportions of differentially expressed genes ($\pi_1 = 1 - \pi_0$) at 0.005, 0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.75, and 0.80 for each simulation study. The vector μ is set as a sequence from 1 to 3 with length equal to $10,000 \times \pi_1$. We use t-statistics for two independent samples for testing differential gene expression between groups. Each simulation include 1,000 iterations. Each procedure is set to control the FDR at the 0.05 level.

Simulation results

Figure 1 illustrates FDR, power, and stability estimates of the three multiple testing procedures for small sample size (5 in each group) when $\rho = 0$. For test statistics that were independent from each other ($\rho = 0$),the Benjamini-Hochberg procedure has the strongest FDR control (the smallest FDR), followed by the Bon-EV procedure. Storey's q-value procedure has the largest FDR, although all three procedures control FDR within 5%. It is also noticeable that the estimated FDR decreases as the proportion of non-true null hypotheses increase for all three multiple testing procedures. Storey's q-value procedure has the greatest power, followed first by the Bon-EV procedure, and then the Benjamini-Hochberg procedure (Additional file 1: Table S1). All three multiple testing procedures have very low power when the proportions of non-true null hypotheses are less than 15%, and the power of all multiple testing procedures increases as proportions of non-true null hypotheses increase. The Benjamini-Hochberg procedure is the most stable, followed by the Bon-EV procedure, with Storey's q-value procedure the least stable. The Benjamini-Hochberg procedure has the smallest SD of total number of discoveries and Storey's q-value procedure has the largest SD of total number of discoveries (Additional file 1: Figure S1). When test statistics are moderately correlated ($\rho = 0.4$), the same trends are observed for FDR, power, and SD of total number of discoveries. We notice the power of our Bon-EV

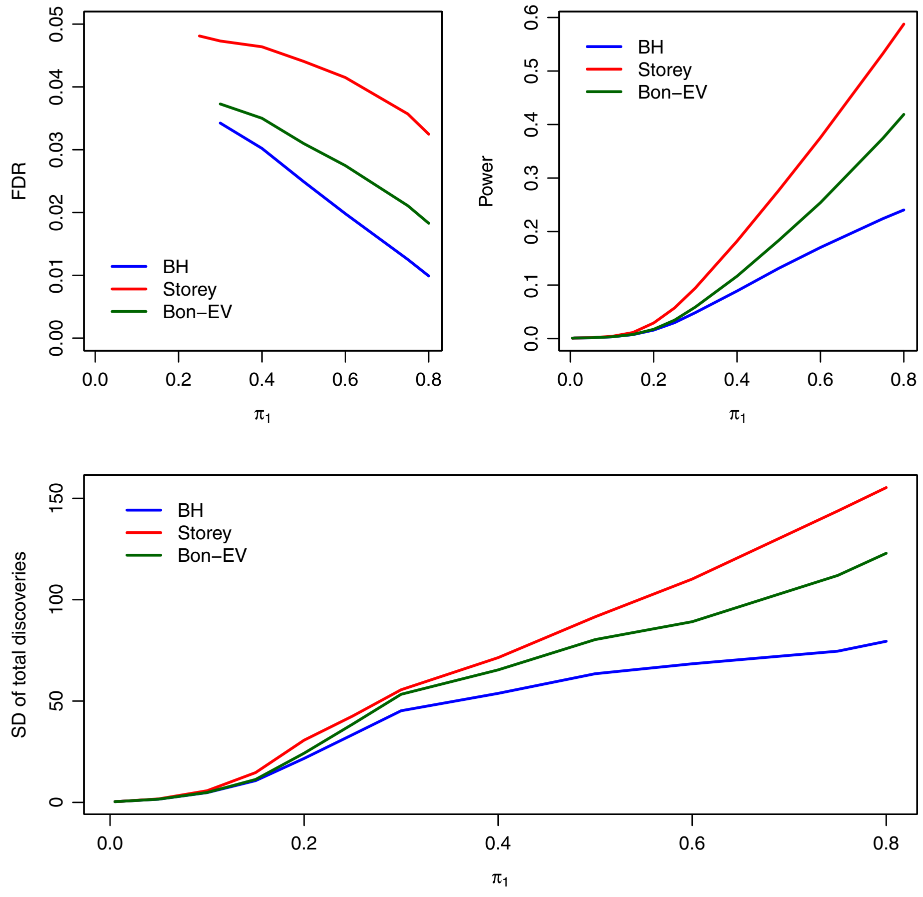

Fig. 1 Estimated FDR, power, SD of power, and SD of total number of discoveries of compared multiple testing procedures with $\rho = 0$ and sample size of 5 in each group. *Blue*: Benjamini-Hochberg procedure; *Red*: Storey's q-value procedure; *Dark Green*: Bon-EV procedure

procedure and Storey's q-value procedure converge as the correlation between test statistics increases from 0 to 0.4. Meanwhile, the power of all three procedures increases as correlations of test statistics increase especially when the proportions of non-true null hypotheses are small. When correlations within test statistics further increase to 0.8, we observe the same trends for FDR, power, and SD of total number of discoveries (Additional file 1: Figure S2). It is also noticeable that the difference in power between Storey's q-value procedure and our Bon-EV procedure gets increasingly smaller as test statistic correlations increase to 0.8 from 0.4. The total number of discoveries (Additional file 1: Table S3) increase as the correlation of test statistics increases. The total number of discoveries of our Bon-EV procedure is close to Storey's q-value procedure and higher than the Benjamini-Hochberg procedure. In these small sample size cases, a large proportion of non-true null hypotheses are not detected, especially when the proportion of non-true null hypotheses is small. The FDR, power, and stability estimates when the correlations across genes are random are close to the estimates when correlations across genes are 0.4 (ρ = 0.4, Additional file 1: Figure S3 and Additional file 1: Table S1).

Figure 2 illustrates the FDR, power, and SD of total number of discoveries of the three multiple testing procedures at sample sizes of 15 in each group when $\rho = 0$. At these sample sizes and with independent test statistics ($\rho = 0$), all three multiple testing procedures control FDR to within 5% except when the proportion of non-true null hypotheses is very small. The power of our Bon-EV procedure is very close to Storey's q-value procedure and higher than the Benjamini-Hochberg procedure when the proportion of non-true null hypotheses is greater than 0.15 (Additional file 1: Table S2). Storey's q-value procedure has the lowest stability (the highest SD of total number of discoveries) among the three multiple testing procedures especially when the proportion of non-true null hypotheses is large. We observe the same trend at a moderate level of correlation between test statistics (ρ = 0.4). The power of our Bon-EV procedure is the same as the power of Storey's q-value procedure, and the Bon-EV procedure continues to show greater stability (smaller SD of total number of discoveries) than Storey's q-value procedure

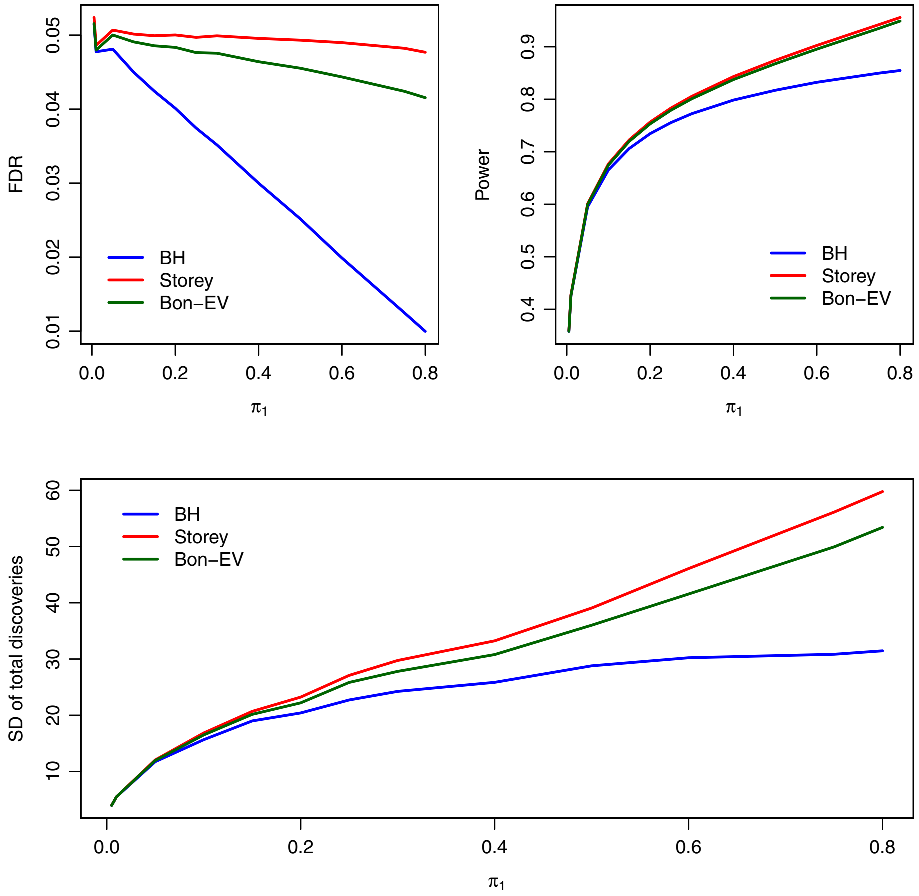

Fig. 2 Estimated FDR, power, SD of power, and SD of total number of discoveries of compared multiple testing procedures with $\rho = 0$ and sample size of 15 in each group. *Blue*: Benjamini-Hochberg procedure; *Red*: Storey's q-value procedure; *Dark Green*: Bon-EV procedure

(Additional file 1: Figure S4). At the highest level of correlation between test statistics ($\rho = 0.8$), the trends remain the same (Additional file 1: Figure S5). The power of our Bon-EV procedure and Storey's q-value procedure is greater than the Benjamini-Hochberg procedure, and the Benjamini-Hochberg procedure has better FDR control and lower SD of total number of discoveries. Our Bon-EV procedure still has better FDR control, the same power, and higher stability than Storey's q-value procedure. The total number of discoveries is much closer to the true number of non-true null hypotheses when the sample size increases to 15 in each group (Additional file 1: Table S3). The total number of discoveries for the Bon-EV procedure is almost the same as for Storey's q-value procedure, and still higher than the Benjamini-Hochberg procedure. Also, for Storey's q-value procedure, the total number of discoveries exceeds the number of non-true null hypotheses when the proportion of non-true null hypotheses is larger than 40%. Similar results on FDR, power, and stability estimates are observed when the data have random correlations (Additional file 1: Figure S6 and Additional file 1: Table S2).

Figure 3 shows the FDR, power, and SD of total number of discoveries of the three multiple testing procedures when $\rho = 0$ and sample size is as large as 30 in each group. All three multiple testing procedures illustrate reasonable control of FDR to within 5%. The power of our Bon-EV procedure is equivalent to Storey's q-value procedure but higher than the Benjamini-Hochberg procedure when the proportion of non-true null hypotheses is greater than 0.15 (Additional file 1: Table S3). Storey's q-value procedure still has the lowest stability among the three multiple testing procedures especially when the proportion of non-true null hypotheses is large. We observe similar results on FDR, power, and stability estimates at a moderate and high level of correlations between test statistics when $\rho = 0.4$ and $\rho = 0.8$ (Additional file 1: Figure S7 and Additional file 1: Figure S8). We also notice that the power and stability improve as the correlation and sample size increase. The FDR, power, and stability estimates when data have random correlations are similar to those estimates when data has moderate correlations ($\rho = 0.4$, Additional file 1: Figure S9 and Additional file 1: Table S3 and Table S4).

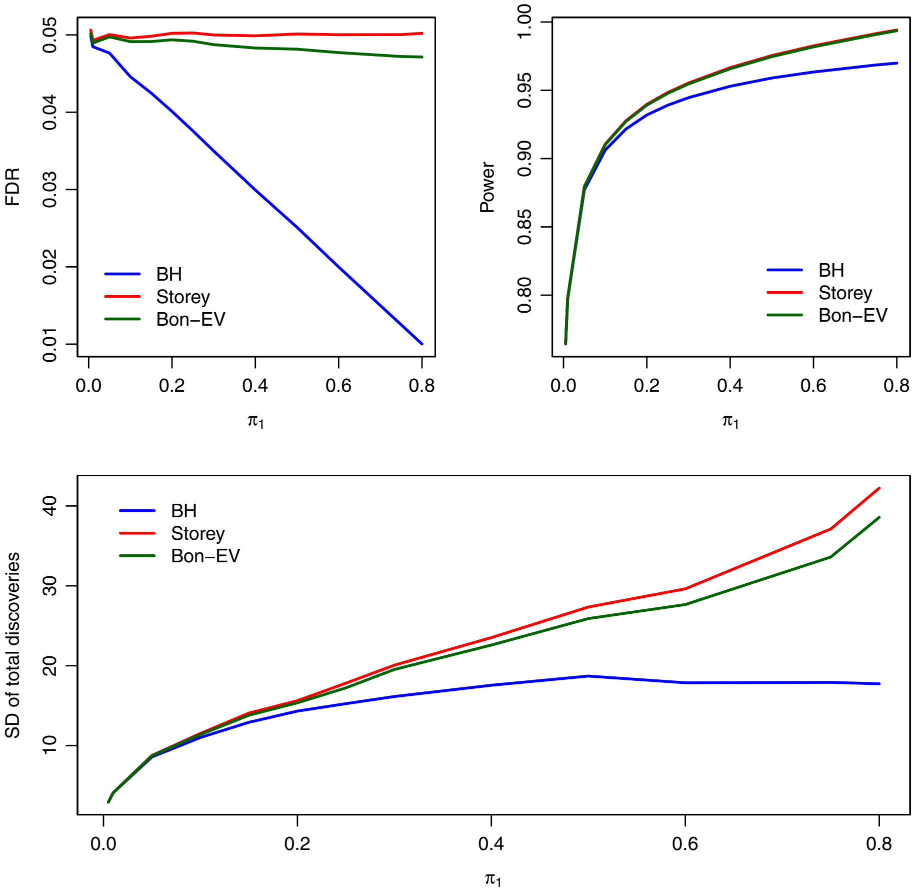

Fig. 3 Estimated FDR, power, SD of power, and SD of total number of discoveries of compared multiple testing procedures with $\rho = 0$ and sample size of 30 in each group. *Blue*: Benjamini-Hochberg procedure; *Red*: Storey's q-value procedure; *Dark Green*: Bon-EV procedure

Examples using real data

As a complement to our simulation studies, we also compare apparent test power (total number of discoveries) and stability (SD of total number of discoveries) resulting from these three different procedures using human RNA-Seq data [12]. We compare gene expression levels between 17 females and 24 males using count data from RNA-Seq downloaded from the ReCount web site [13]. The RNA-Seq data from human B-cells that we analyze include 52,580 genes and 41 samples. The summarized count data from the RNA-Seq experiment is first filtered by only retaining genes that express at a count-per-million (CPM) above 0.5 in at least two samples. The retained 9745 genes are further normalized to eliminate RNA composition biases between libraries by finding a set of scaling factors for the library sizes that minimize the log-fold changes between the samples for most genes using the default method-a trimmed mean of M values (TMM) [14] between each pair of samples. The raw p-value is obtained by fitting negative binomial generalized linear models with Cox-Reid dispersion estimates using the glmFit and flmLRT function in the edgeR package in Bioconductor [15].

We apply the three multiple testing procedures to the raw p-values generated from the edgeR package to compare the total number of rejected genes (apparent test power) after controlling the false discovery rate. Figure 4 shows the apparent test powers of these three different multiple testing procedures at different FDR levels ranging from 0.01 to 0.05. The Bon-EV procedure and Storey's q-value procedure produce the same number of rejections at the FDR level of 0.02 to 0.04. The Bon-EV procedure discovers more genes than the Benjamini-Hochberg procedure and Storey's q-value procedure when the FDR level is equal to 1%. The apparent test power comparison is consistent with our simulation results.

To examine the stability of the Bon-EV procedure, we bootstrap the RNA-Seq samples 1000 times within each group and obtain total number of rejections in each boostrap sample. Then, we examine the stability of the Bon-EV, the Benjamini-Hochberg procedure and the Storey's q-value procedure by calculating the standard deviation of total number of rejections from 1000 bootstrap samples. Figure 5 shows the stability comparison results among these three procedures, which are also consistent with the results from our simulation studies.

Using a similar approach, we compare apparent test power and stability of the three multiple testing procedures using p-values generated from a differential gene expression analysis of the two most commonly used inbred mouse strains in neuroscience research - C57BL/6J (B6) and DBA/2J (D2) [16]. Data from 10 B6 mouse

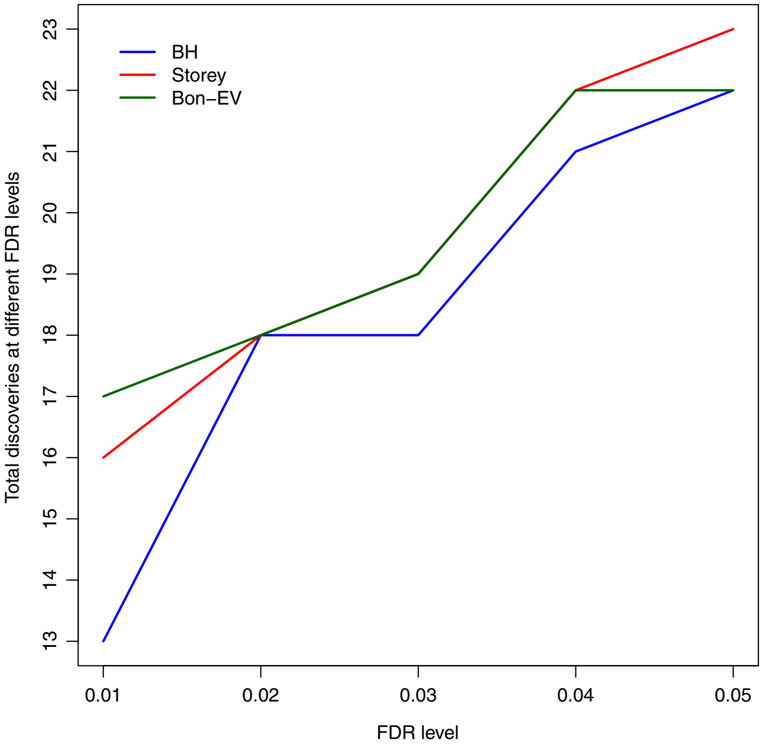

Fig. 4 Total number of discoveries from human RNA-Seq data by three different multiple testing procedures from Cheung data. *Blue*: Benjamini-Hochberg procedure; *Red*: Storey's q-value procedure; *Dark Green*: Bon-EV procedure

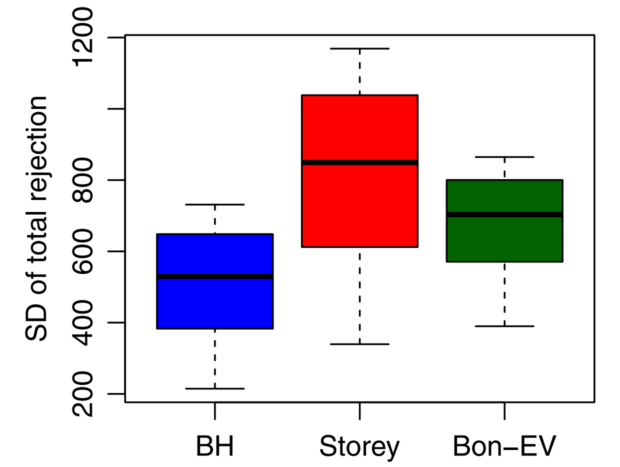

Fig. 5 Stability comparisons among three different multiple testing procedures using Cheung data. *Blue*: Benjamini-Hochberg procedure; *Red*: Storey's *q*-value procedure; *Dark Green*: Bon-EV procedure

samples and 11 D2 mouse samples in the RNA-Seq experiment and the count data are again downloaded from the ReCount web site. Using the same filtering criteria with $CPM > 0.5$ in at least two samples, we retain 11471 genes out of 36536 total genes. After using the same analysis method from the edgeR package in Bioconductor to obtain the raw p-values, we apply the three different multiple testing procedures to calculate adjusted p-values. Figure 6 shows that the number of discoveries at different FDR levels ranges from 0.01 to 0.05. Our Bon-EV procedure also has more discoveries at lower FDR levels. The stability shown in Fig. 7 indicates that the Bon-EV procedure has higher stability than Storey's q-value procedure. Again, the results are consistent with our simulation results.

Discussion

In this study, we propose a new multiple testing procedure (Bon-EV), based on the Bonferroni procedure, intended to improve FDR control and stability as well as maintain power. We compare the Bon-EV procedure with the Benjamini-Hochberg procedure and Storey's q-value procedure using both simulation studies and real RNA-Seq data. Our studies show that the proposed Bon-EV multiple testing procedure has better control of FDR and higher stability than Storey's q-value procedure, and also maintains high levels of power at small and medium sample sizes.

Next generation sequencing and third generation sequencing technology has become more popular in biological and biomedical studies. The sample size in sequencing studies remain small to medium although

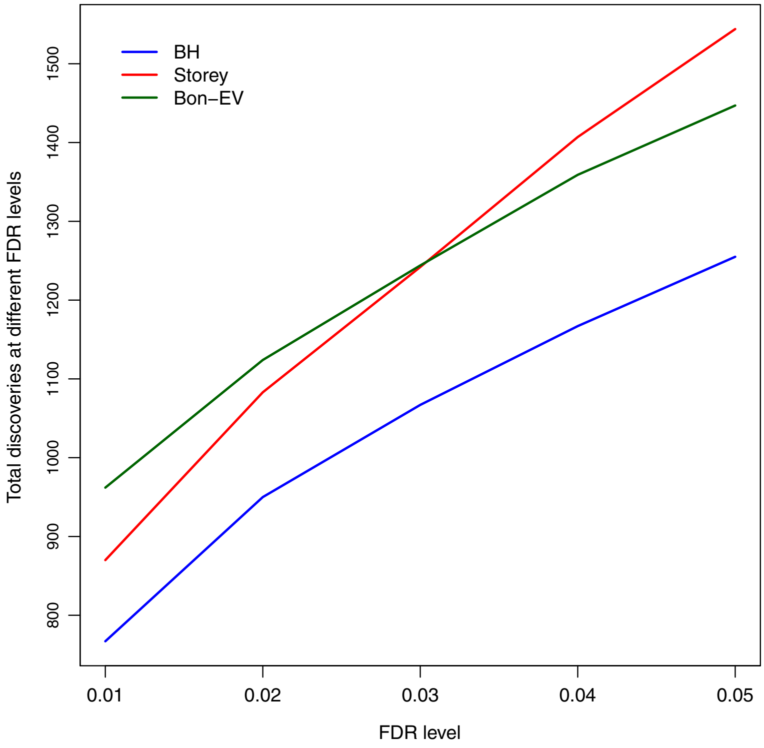

Fig. 6 Total number of discoveries from human RNA-Seq data by three different multiple testing procedures from Bottomly data. *Blue*: Benjamini-Hochberg procedure; *Red*: Storey's *q*-value procedure; *Dark Green*: Bon-EV procedure

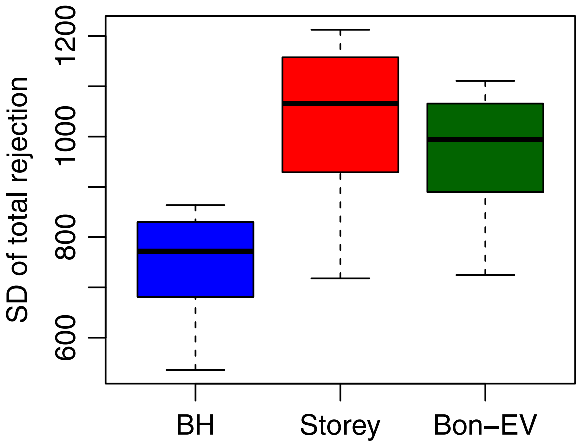

Fig. 7 Stability comparisons among three different multiple testing procedures using Bottomly data. *Blue*: Benjamini-Hochberg procedure; *Red*: Storey's *q*-value procedure; *Dark Green*: Bon-EV procedure

the price of sequencing per sample has significantly decreased in recent years. Multiple testing procedures with larger power could help increase the probability of novel discoveries. The Bon-EV procedure, similar to the Storey's *q*-value procedure, offers higher power than the Benjamini-Hochberg procedure and increases the cost-effectiveness of sequencing studies.

Compared to the Storey's *q*-value procedure, the Bon-EV procedure offers better FDR control. In recent years, irreproducibility of results in biomedical research has drawn increasing attention in the popular press and academia [17, 18]. Investigations have found many landmarks in preclinical oncology to be non-reproducible, and confirming research conducted by scientists in the haematology and oncology department at the biotechnology firm Amgen in Thousand Oaks, California, finds only 11% of scientific findings they examined to be reproducible [17]. The reasons for irreproducibility are at least, in part, due to false discoveries in those studies [18–20]. With better FDR control, the Bon-EV procedure will enable more accurate control of false discoveries in genomic studies. Meanwhile, understanding the source of variation in the data generation and analysis can help improve reproducibility of scientific studies [21]. Multiple testing procedures are widely used to select significant features such as genes, SNPs, methylation loci, and others in microarray and NGS studies. Assessment of the stability of statistical findings from multiple testing procedures and improving the stability of these procedures could reduce replication failures. The Bon-EV procedure will help reduce replication failures compared with the Storey's *q*-value procedure, and provide higher stability.

Conclusions

Our study investigates the stability of Benjamini-Hochberg and Storey's *q*-value FDR controlling procedures commonly used in genomic and genetic data analysis and proposes a new multiple testing procedure with higher stability, better FDR control and power equivalent to Storey's *q*-value procedure as well as higher power than the Benjamini-Hochberg procedure. The Bon-EV multiple testing procedure we propose is attractive in microarray and sequencing data analysis in that it has higher power than the Benjamini-Hochberg procedure, and better FDR control and higher stability than Storey's *q*-value procedure.

Additional file

Additional file 1: The additional file includes supplemental figures when correlations were 0.4, 0.8, and random across all genes. Figures S1-S9 shows the estimated FDR, power, SD of power, and SD of total discoveries of compared multiple testing procedures, with correlations of 0.4, 0.8, and random, and sample size of 5, 15, and 30 in each group. The values of power and stability were shown in supplemental Table S1-S4. **Figure S1.** Shows the performance of compared multiple testing procedures with ρ = 0.4 and sample size of 5 in each group. **Figure S2.** Shows the performance of compared multiple testing procedures with ρ= 0.8 and sample size of 5 in each group. **Figure S3.** Shows the performance of compared multiple testing procedures with random correlation across genes and sample size of 5 in each group. **Figure S4.** Shows the performance of compared multiple testing procedures with ρ = 0.4 and sample size of 15 in each group. **Figure S5.** Shows the performance of compared multiple testing procedures with ρ = 0.8 and sample size of 15 in each group. **Figure S6.** Shows the performance of compared multiple testing procedures with random correlation across genes and sample size of 15 in each group. **Figure S7.** Shows the performance of compared multiple testing procedures with ρ = 0.4 and sample size of 30 in each group. **Figure S8.** Shows the performance of compared multiple testing procedures with ρ = 0.8 and sample size of 30 in each group. **Figure S9.** Shows the performance of compared multiple testing procedures with random correlation across genes and sample size of 30 in each group. **Table S1.** Shows the power and stability of compared multiple testing procedures for sample size n = 5 in each group. **Table S2.** Shows the power and stability of compared multiple testing procedures for sample size n = 15 in each group. **Table S3.** Shows the power and stability of compared multiple testing procedures for sample size n = 30 in each group. **Table S4.** Shows the total number of rejections of compared multiple testing procedures for sample size n = 5, 15, 30 in each group. (PDF 59 kb)

Acknowledgements

We thank Stanley Pounds from biostatistics department at St. Jude Children's Research Hospital for his insightful suggestions and comments on our manuscript and the Center for Integrated Research Computing at the University of Rochester for providing high performance computing resources. We also would like to thank the anonymous reviewers for their insightful comments and suggestions that helped to further improve our manuscript.

Funding

Dr. Li's, Dr. Zand's, Mr. Fogg's, and Dr. Dye's time is supported by the University of Rochester's Clinical and Translational Science Award (CTSA) number UL1 TR000042 and UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health. Dr. Zand is also supported by the National Institute of Allergy and Infectious Diseases and the National Institute of Immunology, grant numbers AI098112 and AI069351. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials

The Bon-EV multiple testing procedure is available as the BonEV package in R for download at https://CRAN.R-project.org/package=BonEV. The RNA-Seq data used in the real data examples are downloaded from http://bowtie-bio.sourceforge.net/recount/ website.

Authors' contributions

DL, ZX, MZ, TF, and TD contributed to the conception and design of the study. DL, ZX designed the simulation studies. DL performed the simulation studies and real data analysis. DL, ZX, MZ, TF, and TD interpreted the analysis results and wrote the manuscript. All authors edited and approved the final manuscript. All authors agreed to be accountable for all aspects of the study in ensuring that questions related to the accuracy or integrity of any part of the study are appropriately investigated and resolved.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable

Ethics approval and consent to participate

Not applicable

Author details

[1]Clinical and Translational Science Institute, School of Medicine and Dentistry, University of Rochester, 265 Crittenden Boulevard CU 420708, 14642 Rochester, NY, USA. [2]Goergen Institute for Data Science, University of Rochester, Computer Studies Building, 14642 Rochester, NY, USA. [3]Department of Obstetrics and Gynecology, University of Rochester, 500 Red Creek Drive Suite 220, 14623 Rochester, NY, USA.

Received: 20 July 2016 Accepted: 7 December 2016
Published online: 03 January 2017

References

1. Pounds SB. Estimation and control of multiple testing error rates for microarray studies. Brief Bioinforma. 2006;7(1):25–36.
2. Hochberg Y, Tamhane AC. Multiple Comparison Procedures. New York, NY: Wiley-Interscience; 1987.
3. Dohler S. A sufficient criterion for control of some generalized error rates in multiple testing. Stat Probab Lett. 2014;92:114–20.
4. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B (Methodological). 1995;57(1):289–300.
5. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. J Educ Behav Stat. 2000;25(1):60–83.
6. Keselman HJ, Cribbie R, Holland B. Controlling the rate of type i error over a large set of statistical tests. Br J Math Stat Psychol. 2002;55:27–39.
7. Storey JD. A direct approach to false discovery rates. J R Stat Soc Series B (Methodological). 2002;64(3):479–98.
8. Storey JD. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat. 2003;31(6):2013–35.
9. Khatree R, Naik D. Computational Methods in Biomedical Research. New York, NY: Chapman & Hall/CRC; 2008.
10. Gordon A, Glazko G, Qiu X, Yakovlev A. Control of the mean number of false discoveries, bonferroni and stability of multiple testing. Ann Appl Stat. 2007;1(1):179–90.
11. Li D, Xie Z, Le Pape M, Dye T. An evaluation of statistical methods for dna methylation microarray data analysis. BMC Bioinforma. 2015;16(1): 1–20. doi:10.1186/s12859-015-0641-x.
12. Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, Morley M, Spielman RS. Polymorphic *Cis*- and *Trans*-regulation of human gene expression. PLoS Biol. 2010;8(9):1000480. doi:10.1371/journal.pbio.1000480.
13. Frazee AC, Langmead B, Leek JT. Recount: A multi-experiment resource of analysis-ready rna-seq gene count datasets. BMC Bioinformatics. 2011;12(1):1–5. doi:10.1186/1471-2105-12-449.
14. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of rna-seq data. Genome Biol. 2010;11(3):25.
15. Robinson MD, McCatyhy DJ, Smyth GK. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
16. Bottomly D, Walter NAR, Hunter JE, Darakjian P, Kawane S, Buck KJ, Searles RP, Mooney M, McWeeney SK, Hitzemann R. Evaluating gene expression in c57bl/6j and dba/2j mouse striatum using rna-seq and microarrays. PLoS ONE. 2011;6(3):1–8. doi:10.1371/journal.pone.0017820.
17. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012;483:531–3.
18. Plant AL, Locascio LE, May WE, Gallagher PD. Improved reproducibility by assuring confidence in measurements in biomedical research. Nat Methods. 2014;11(9):895–9.
19. Begley CG. Reproducibility: Six red flags for suspect work. Nature. 2013;497:433–4.
20. McNutt M. Reproducibility. Science. 2014;343(6168):229–9. doi:10.1126/science.1250475. http://science.sciencemag.org/content/343/6168/229.full.pdf.
21. Stodden V. Reproducing statistical results. Annu Rev Stat Appl. 2015;2: 1–19.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central