BIOINFORMATICS DISCOVERY NOTE Vol. 24 no. 10 2008, pages 1225–1228
doi:10.1093/bioinformatics/btn120

Data and text mining

A note on the false discovery rate and inconsistent comparisons between experiments

Roger Higdon[1], Gerald van Belle[2] and Eugene Kolker[1,3,*]

[1]Seattle Children's Research Institute, Seattle, WA 98101, [2]Departments of Biostatistics and Environmental and Occupational Health Sciences, University of Washington and [3]Division of Biomedical Informatics, Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, WA 98195, USA

Received on January 18, 2008; revised on March 14, 2008; accepted on April 1, 2008

Advance Access publication April 19, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The false discovery rate (FDR) has been widely adopted to address the multiple comparisons issue in high-throughput experiments such as microarray gene-expression studies. However, while the FDR is quite useful as an approach to limit false discoveries within a single experiment, like other multiple comparison corrections it may be an inappropriate way to compare results across experiments. This article uses several examples based on gene-expression data to demonstrate the potential misinterpretations that can arise from using FDR to compare across experiments. Researchers should be aware of these pitfalls and wary of using FDR to compare experimental results. FDR should be augmented with other measures such as *p*-values and expression ratios. It is worth including standard error and variance information for meta-analyses and, if possible, the raw data for re-analyses. This is especially important for high-throughput studies because data are often re-used for different objectives, including comparing common elements across many experiments. No single error rate or data summary may be appropriate for all of the different objectives.

Contact: Eugene.Kolker@seattlechildrens.org

1 INTRODUCTION

In an effort to control for multiple comparisons and increase power over more conventional methods (Dudoit *et al.*, 2003), the false discovery rate (FDR) (Benjamini and Hochberg, 1995) has become increasingly popular for large, exploratory data analyses. In particular, the FDR has become the standard criterion for assessing results in microarray gene-expression studies, along with the associated *q*-value (FDR for a specific *p*-value threshold) used to quantify individual comparisons (Allison *et al.*, 2006; Kerr and Churchill, 2001; Storey and Tibshirani, 2003; Tusher *et al.*, 2001). These quantities are defined in Table 1. The FDR is significant in other fields as well, including for example, imaging (Srikanth *et al.*, 2006), proteomics (Karp *et al.*, 2007) and genetic association and linkage (Chen and Storey, 2006).

However, use of the FDR and its associated *q*-value may result in inconsistent and misleading interpretation of comparisons across different experiments. This inconsistency is inherent to other stepwise multiple comparison procedures such as Student–Newman–Keuls (Keuls, 1952) and the Holm Bonferoni adjustment (Holm, 1979). This difficulty is in part due to the omnibus nature of such tests, where many different elements of the tests and family of comparisons can lead to the same error rate. The rapid increase in popularity of the FDR has made it more necessary than ever to demonstrate these inconsistencies. These inconsistencies are fundamental to the FDR and not to issues of estimation for the FDR, a topic which has been discussed at great length elsewhere (Allison *et al.*, 2006; Benjamini and Hochberg, 1995; Storey, 2002; Tsai *et al.*, 2003).

This topic has not been directly addressed in the multitudes of papers discussing the FDR. Few papers demonstrate the potential for interpretation error or issues with comparing the FDR and *q*-values across different experiments. Others have noted potential inconsistencies in the interpretation of any results that used any multiple comparison procedures (O'Brien, 1983; Rothman, 1990). This inconsistency is often due to differences in the number of comparisons as illustrated by the following. Assume that there are two studies, the first compares all pairs of treatments A, B and C, while the other compares only treatments A and B. Focusing on the comparison between A and B, assume both studies observe the same unadjusted *p*-value of 0.03 for the comparison. Using a Bonferoni correction, the first study can adjust for multiple comparisons yielding an adjusted *p*-value of 0.09 (3*0.03) for this comparison. As a result, the studies would reach different conclusions based on a standard 0.05 *p*-value threshold, despite observing the same difference between A and B.

Methods to control for the FDR are relatively less sensitive to the number of comparisons than other procedures that adjust for multiple comparisons (Holland and Cheung, 2002). However, even when the numbers of comparisons are identical across different experiments, the thresholds to control for the FDR and the associated *q*-values for individual comparisons

*To whom correspondence should be addressed.

Table 1. Definitions of error rates in a multiple testing situation using the notation of Benjamini and Hochberg (1995)

	Fail to reject H_0	Reject H_0	All decisions
H_0 true	U	V	m_0
H_0 false	T	S	m-m_0
All H_0	m-R	R	m

Significance $= \alpha = V/m_0$.
p-value (pv) is the smallest α for which H_0 rejected.
FDR $= V/R$.
q-value (qv) = FDR for a given p-value threshold.
Power $= 1-\beta = S/(m-m_0)$ (Note, this is different from the power of an individual test, which depends on the alternative hypothesis for each individual test).
Proportion of true null hypotheses $= p = m_0/m$.
Local FDR is the FDR of comparisons on the rejection boundary or equal to a given.

are highly dependent on the results of other comparisons. This situation is often encountered in microarray studies, where many series of experiments are based on the same set of genes.

2 RESULTS AND DISCUSSION

2.1 Ten-gene comparisons

For simplicity, assume there are two studies comparing gene-expression levels between two conditions (expression ratios) for the identical set 10 genes (clearly, this example can be scaled to any number of genes). Focusing on gene X, assume both studies observe the same unadjusted p-value of 0.01 in a test for differential gene expression (expression ratio not equal to one). Assume also that this is the smallest p-value among the 10 genes in the first study and the 3rd smallest in the second study. In the first study a conservative estimate of the FDR using a p-value threshold of 0.01 would be 10% (10*0.01/1) and in the second the FDR at that same threshold would be ~3% (10*0.01/3). Based on a 5% FDR threshold, gene X would be considered differentially expressed in the second study but not in the first. This result appears to be counter intuitive; despite observing the same level of differential expression in gene X, it is considered significant when it is the third smallest p-value, but is no longer significant when it is the smallest.

As we shall illustrate in the rest of this article, the nature of the FDR is such that the larger the pool of differentially expressed genes, the less conservative the p-value threshold becomes. This result makes sense probabilistically, however, practically and intuitively should the significance of this p-value change in this way? Common sense might suggest the opposite. One should consider larger p-values when there are only few differentially expressed genes, not when there are many.

2.2 FDR dependencies

The following equation describes the relationship between the q-value (qv) and the p-value (pv) of an individual comparison. It shows how the q-value is dependent on the totality of comparisons in an experiment.

$$qv = \frac{V}{S+V} = \frac{pv \cdot m_0}{(1-\beta)\cdot(m-m_0)+pv\cdot m_0} = \frac{pv \cdot p}{(1-\beta)\cdot(1-p)+pv\cdot p}. \quad (1)$$

Notation is defined in Table 1 following Benjamini and Hochberg (1995). Note that the power $(1-\beta)$ depends on the significance level or p-value threshold, the particular statistical test, as well as the distribution of alternative hypotheses. For a specific FDR (i.e. 5%), the p-value threshold will therefore depend upon the overall power and the proportion of true null hypotheses (i.e. the proportion of equally expressed genes).

Figure 1 shows the p-values corresponding to an FDR or q-value of 5% at different proportions of true null hypotheses and power. The p-value threshold decreases as power decreases and the proportion of null hypotheses (equally expressed genes) increases. At the upper left corner of the figure, where large numbers of hypotheses are rejected (power = 0.9 and 50% null hypotheses), the p-value threshold to achieve a 5% FDR is near 0.05, which is on the border of what would be considered significant for a single comparison. On the other hand, at the bottom right corner, where few hypotheses were rejected (power = 0.25 and 95% null hypotheses), the p-value threshold is below 0.001, a highly significant difference for a single comparison.

The local FDR can be defined as the FDR for genes equal to given q-value (or p-value), where the q-value is the FDR for genes with p-values as small or smaller (Table 1). It has been argued that the FDR can be misleading because the error rate on the rejection boundary (local FDR) is often much higher than the overall FDR (Efron, 2004). Therefore, using the local FDR to judge significance might be preferable. However, the results shown in Figure 1 are not unique to the overall FDR, but similarly affect the local FDR. For example, see Figure 2 where the distribution of the test statistics (i.e. based on log-expression ratios) is assumed to be $N(\mu,1)$ where μ is the true log-expression ratio and $\mu = 0$ under the null hypothesis and the distribution of μ under the alternative hypothesis is $N(1,1)$. This formulation allows easy calculation of p-values, FDR and the local FDR (Efron, 2004). This results in a similar relationship of p-value thresholds (to achieve 5% FDR) with the proportion of true null hypotheses as was shown in Figure 1. However, while the local FDR was higher than the overall 5% FDR, it does not vary much with the proportion of true null hypotheses used in Figure 2, ranging from 11.8% down to 10.1%. In fact, if we hold the local FDR fixed at 10%, the curve changes only slightly from the fixed 5% FDR curve as shown in Figure 2, where similar variation in p-value thresholds is apparent. This demonstrates that the local FDR suffers from the same difficulties and issues as the FDR.

2.3 Mouse liver experiments

The following example is based on microarray data obtained from the Gene Expression Omnibus (GEO) data repository developed by the National Center for Biotechnology Information at the NIH (Barrett *et al.*, 2007). The study compared mouse liver samples for a treatment (PPARα agonist Wy14643)

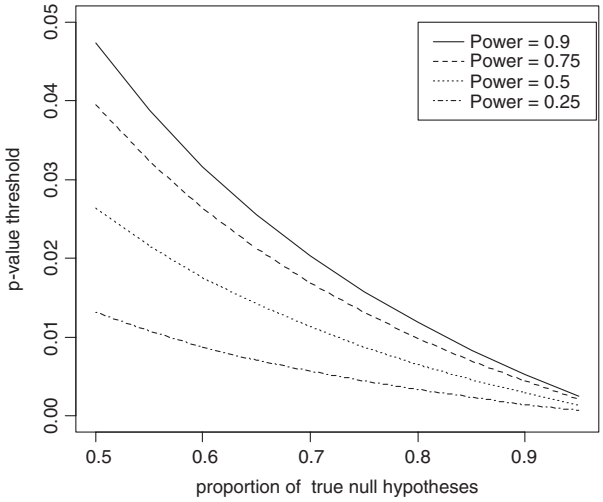

Fig. 1. Variation in *p*-value threshold for a fixed FDR of 5%. Plot shows *p*-value threshold to achieve a 5% FDR (or *p*-value corresponding to a *q*-value of 0.05) as power and the proportion of true null hypotheses vary.

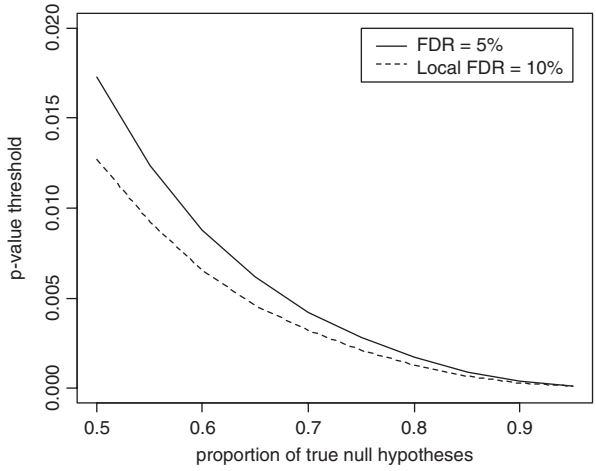

Fig. 2. Similarity in variation in *p*-value threshold for a fixed 5% FDR and 10% local FDR. Plot shows *p*-value threshold to achieve a 5% FDR and 10% local FDR as a function of the proportion of true null hypotheses. The distribution of the test statistics (i.e. based on log-expression ratios) is assumed to be $N(\mu,1)$ where μ is the true log-expression ratio and $\mu = 0$ under the null hypothesis and the distribution of μ under the alternative hypothesis is $N(1,1)$.

Table 2. Comparison of *q*-values, *p*-values and expression ratios (ER) for three genes from two different mouse liver microarray experiments (GEO experiment series GSE8295)

Gene	Wild-type experiment					Mutant experiment				
	q-value	*p*-value	ER	*p*-val. rank	ER rank	*q*-value	*p*-value	ER	*p*-val. rank	ER rank
Per3	0.03	0.007	1.76	7656	3522	0.06	0.00003	2.75	17	44
Hlf	0.02	0.003	1.53	6509	5433	0.08	0.00005	1.58	25	556
Arntl	0.02	0.003	0.14	6372	316	0.09	0.00009	0.12	38	6

The ranking of the gene in terms of *p*-value (or equivalently *q*-value) and ER is also given.

versus a control (GEO experiment series GSE8295). The experiment was repeated for wild-type and mutant mice. The array contains ~40 000 sequences (genes and variants) and there were four replicates in each treatment group. Log-expression values were analyzed using the Limma package for the R programming language (Smyth, 2003) to generate *p*-values and *q*-values for differential expression studies.

Analysis of the two experiments (wild-type and mutant) results in disproportionate numbers of differentially expressed genes at a 5% FDR: 8669 for the wild-type and only 16 for the mutant. In turn, this results in dramatically different *p*-value thresholds to achieve a 5% FDR (0.014 versus 0.00002), mirroring the example shown in Figure 1. Using 5% FDR (*q*-value <0.05) as the threshold, there are a number of significant, differentially expressed genes in the wild-type experiment with larger *p*-values, smaller expression ratios and much lower rankings than non-significant genes in the mutant experiment. The results for three such genes are shown in Table 2.

Clearly, a simple comparison of *q*-values between the two experiments can be quite misleading for specific genes. If a researcher was specifically interested in these three genes and only had a list of differentially expressed genes with a 5% FDR, the wrong conclusion is inevitable: these three genes are significantly differentially expressed in the wild-type experiment but not in the mutant experiment.

2.4 Use of FDR

FDR is a useful concept and control for multiple testing issues, particularly for the huge number of comparisons made in high-throughput experiments such as microarray gene-expression studies. However, relying only on the FDR to judge the significance of results across different experiments can lead to inconsistencies and misinterpretation of individual comparisons.

The FDR is an appropriate error measure to identify a list of genes that has a suitable high likelihood of being differentially expressed based only on the information from the specific experiment. It may be advisable not to use a pre-set FDR threshold, since in some circumstances it may result in huge numbers of candidate genes, while in others it may yield only a few.

The FDR is only one type of useful information for evaluating individual comparisons. For instance, considering the per comparison error rate (p-value), the magnitude of the difference (i.e. expression ratios) and perhaps the local FDR will give a more complete picture of the significance of different genes. The most appropriate criteria depend upon the objective(s) of the study. For example, if one wants to ensure each individual gene has a high likelihood of differential expression, then the local FDR is more appropriate. If one wants to rank the most differentially expressed genes then, as was seen in the mouse liver experiment (Table 2), the FDR, local FDR and p-value all result in the same ranking of genes, while a ranking based on the expression ratios is quite different.

When a researcher's interest is in examining a single gene or a small group of genes across different experiments, the FDR, q-values and local FDR are not the appropriate measures. The research question now focuses on cross-experimental comparisons, rather than using a single, entire experiment as the basis for analysis. In this case, using the FDR, q-value or local FDR may lead one to exclude comparisons or genes that show consistently small p-values and large differences, but did not achieve a desired FDR in all those studies. Therefore, the FDR, q-values or local FDR can give the false impression that across studies results were inconsistent; p-values and expression ratios will be far more informative for comparing the same small set of genes across different experiments.

A meta-analysis of the experiments (Choi *et al*., 2003) or, better still, a re-analysis of the raw data on the restricted set of genes may provide error rates more specific to this situation. For example, tests for a difference in expression ratios between the mutant and wild-type experiments based on a combined analysis for the three genes in Table 2 show no difference for *Hlf* and *Arnt*l (both p-values $= 0.85$). It is also suggestive that the expression ratio for *Per3* was larger for the mutant experiment (p-value $= 0.07$). These results are contradictory to the results based on the FDR or q-values.

3 CONCLUSION

Gene expression and other large-scale analyses may initially have the objective of finding biomarkers or other discovery targets, and with such an objective, using the FDR is a sensible method for controlling errors and maximizing the number of potential discoveries. However, the data from these studies may serve many purposes, including much more specialized and targeted analyses. It is important that the reported results from these studies include more than simple gene lists for a given FDR.

Including results for all genes and additional quantitative information such as p-values, expression ratios and local FDR values can help researchers make better comparisons across different experiments. Reporting information on variability, standard errors and sample size may make meta-analyses possible. Better still is to make raw data available along with detailed information about the experimental design and data normalization, such as that which is being done by the NIH with GEO. This will allow researchers to estimate the appropriate error rate for the objectives of the study and avoid inconsistent comparisons that obscure scientific discovery.

ACKNOWLEDGEMENTS

We greatly appreciate Caroline Dombrowski, Katie Kerr and Jared Roach for their insightful comments.

Funding: This work was supported by the grants from the National Institutes of Health, National Institute of General Medical Sciences (Grant No. GM076680-01A1) and from the National Science Foundation, Offices of Biological Infrastructure and Molecular and Cellular Biology (Grant No. 0544757) to E.K.

Conflict of Interest: none declared.

REFERENCES

Allison,D.B. *et al*. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet*., **7**, 55–65.

Barrett,T. *et al*. (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*., **35**, D760–D765.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Ser. B*., **57**, 289–200.

Chen,L. and Storey,J.D. (2006) Relaxed significance criteria for linkage analysis. *Genetics*, **173**, 2371–2381.

Choi,J.K. *et al*. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19** (Suppl. 1), i84–i90.

Dudoit,S. *et al*. (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci*., **18**, 71–103.

Efron,B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Stat. Assoc*., **99**, 96–104.

Holland,B. and Cheung,S.H. (2002) Familywise robustness criteria for multiple-comparison procedures. *J. Royal Stat. Soc. Ser. B*., **64**, 63–77.

Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat*., **6**, 65–70.

Karp,N.A. *et al*. (2007) Experimental and statistical considerations to avoid false conclusions in proteomic studies using differential in-gel electrophoresis. *Mol. Cell Proteomics*, **8**, 1354–1364.

Kerr,M.K. and Churchill,G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183–201.

Keuls,M. (1952) The use of the studentize range in connection with an analysis of variance. *Euphytica*, **1**, 112–122.

O'Brien,P.C. (1983) The appropriateness of analysis of variance and multiple-comparison procedures. *Biometrics*, **39**, 787–794.

Rothman,K.J. (1990) No adjustments are needed for multiple comparisons. *Epidemiology*, **1**, 43–46.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*., **3**, Article 3, 1–24.

Srikanth,R. *et al*. (2006) Estimation of false discovery rates for wavelet-denoised statistical parametric maps. *Neuroimage*, **33**, 72–84.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. Royal Stat. Soc. Ser. B*., **64**, 479–498.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Tsai,C.A. *et al*. (2003) Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics*, **59**, 1071–1081.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.