OIKOS 108: 643–647, 2005

Implementing false discovery rate control: increasing your power

Koen J. F. Verhoeven, Katy L. Simonsen and Lauren M. McIntyre

Verhoeven, K. J. F., Simonsen, K. L. and McIntyre, L. M. 2005. Implementing false discovery rate control: increasing your power. – Oikos 108: 643–647.

Popular procedures to control the chance of making type I errors when multiple statistical tests are performed come at a high cost: a reduction in power. As the number of tests increases, power for an individual test may become unacceptably low. This is a consequence of minimizing the chance of making even a single type I error, which is the aim of, for instance, the Bonferroni and sequential Bonferroni procedures. An alternative approach, control of the false discovery rate (FDR), has recently been advocated for ecological studies. This approach aims at controlling the proportion of significant results that are in fact type I errors. Keeping the proportion of type I errors low among all significant results is a sensible, powerful, and easy-to-interpret way of addressing the multiple testing issue. To encourage practical use of the approach, in this note we illustrate how the proposed procedure works, we compare it to more traditional methods that control the familywise error rate, and we discuss some recent useful developments in FDR control.

K. J. F. Verhoeven and L. M. McIntyre, Computational Genomics, Dept of Agronomy, Purdue Univ., Lilly Hall of Life Sciences, 915 W. State Street, West Lafayette, IN 47907-2054. USA (kverhoeven@purdue.edu). – K. L. Simonson, Dept of Statistics, Purdue Univ., 150 N. University Ave. West Lafayette, IN 47907-2068, USA.

The problem

The appropriate threshold to declare a test statistic's p value significant becomes complex when more than one test is performed. In the absence of a true effect each test has a chance of α to yield a significant result, and the chance of drawing at least one false conclusion increases rapidly with the number of tests performed. Protection against false rejections of the null hypothesis, or type I errors, is usually achieved via a Bonferroni-type correction procedure (Holm 1979). By performing individual tests at error rates that are a fraction of the overall nominal α, the chance of making even a single type I error can be maintained at the desired α level (usually 5%). This is called control of the familywise error rate (FWER). With an increasing number of tests, maintaining a low chance of making even one type I error comes at the direct cost of making more type II errors, i.e. not recognizing a true effect as significant. The classical Bonferroni procedure, which performs each of m tests at a type I error rate of α/m, is undesirable because of this trade-off: only a very strong effect is likely to be recognized as significant when many tests are performed. Several improvements to the classical Bonferroni have been proposed in order to reduce the problem of low power (reviewed by García 2004). For instance, the well known Holm's step-down or sequential Bonferroni procedure (Holm 1979, popularized among evolutionary biologists and ecologists by Rice 1989) performs tests in order of increasing p values, and conditional on having rejected tests with smaller p values an increasingly permissive threshold can be used while maintaining the FWER at the desired level (5%). Further power gains are possible with the sequential approach by using a step-up instead of a step-down procedure (that is, testing in order of decreasing p values, Hochberg 1988), by estimating the number of true null hypotheses and correcting for those instead of all tests performed

Accepted 7 September 2004

ISSN 0030-1299

(Schweder and Spjøtvoll 1982, Hochberg and Benjamini 1990), and by accounting for correlations within the dataset which can reduce the effective number of independent tests performed (Cheverud 2001). Although an improvement over the classical Bonferroni, all these procedures still focus on limiting the chance of making even a single type I error, and as such will result in more type II errors than is perhaps desired.

The problem with this approach to type I error control, which has led some to suggest that we should abandon correcting for multiple testing altogether (Moran 2003), is that overall interpretation may not be erroneous if one test is falsely rejected, but may be severely affected by a large number of type II errors. FWER control offers limited opportunity to strike a sensible compromise between the two types of error. The α level can be raised to a 10% (or even 20% or 50%) chance of making at least one type I error, thereby decreasing the rate of type II errors. But with a growing number of tests this still leads to an increasing number of type II errors; it is inherent to controlling the chance of making even a single type I error.

Controlling the false discovery rate

An elegant way to deal with the problem, that was recently advocated for ecological studies by García (2003, 2004), is to control the proportion of significant results that are in fact type I errors ('false discoveries') instead of controlling the chance of making even a single type I error. This new approach was developed by Benjamini and Hochberg (1995). To see the difference between FWER and the false discovery rate (FDR), consider the potential outcomes of each test (Table 1). FWER is the probability that V, the number of type I errors, is greater than or equal to one. FDR, as defined by Benjamini and Hochberg (1995), is the expected proportion of type I errors among all significant results (V/r). Control of FWER, for instance via Bonferroni or sequential Bonferroni adjustment of the per comparison error rate, means that the probability that $V \geq 1$ is maintained at a desired level. Control of FDR means that the expected proportion V/r is maintained at a desired level. When all null hypotheses are true, controlling FWER and FDR are equivalent. In that case either $V/r = 0$ (by definition if $V = 0$) or $V/r = 1$ (if $V > 0$, because all significant results are false), and the expected ratio equals the chance that any false rejection is made. However, if some of the alternative hypotheses are true and $S > 0$, then V/r is either 0 (if $V = 0$) or $0 < V/r < 1$ (if $V > 0$), and the expected ratio is smaller than the chance that any false rejection is made (Benjamini and Hochberg 1995). In those cases FDR is smaller than FWER, and controlling FDR at, say, 5% can result in fewer type II errors than controlling FWER at 5%. The gain increases when more alternative hypotheses are true.

Table 1. Possible outcomes of individual tests. Note that V is the number of type I errors; T is the number of type II errors; and only m, r and $m - r$ are observed while the other variables are unknown.

Truth	Decision		Total
	Not significant	Significant	
Null hypothesis	U	V	m_0
Alternative hypothesis	T	S	$m - m_0$
Total	$m - r$	r	m

The following simple procedure to control FDR at level α was proposed by Benjamini and Hochberg (1995): For m tests, rank the p values in ascending order $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$; Let k be the largest i for which

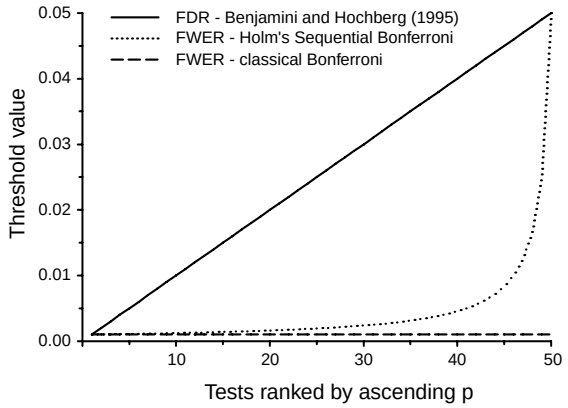
$$P_{(i)} \leq \frac{\alpha}{m} i$$

and reject all null hypotheses $H_{(1)} \ldots H_{(k)}$. In other words, starting with the highest p value each p is checked for this requirement; at the first p that meets the requirement its corresponding null hypothesis and all those having smaller p's are rejected. To visualize the potential for reduction in type II errors using this 1995 Benjamini and Hochberg FDR procedure compared to (sequential) Bonferroni FWER control at the same 5% level, consider the example in Fig. 1. The cost of not wanting to make even a single type I error is reflected in the difference between FWER and FDR significance thresholds for individual tests. Note that if the FDR threshold of 5% yields a list of, for instance, 20 significant results then the expected number of type I errors among these is one.

The above procedure was shown by Benjamini and Hochberg (1995) to control FDR in the case when all m tests are independent, and was also shown (Benjamini and Yekutieli 2001) to control FDR when tests are positively correlated. These are thought to be the most common situations in genetic and ecological studies. Positive dependence, for instance, can occur in marker-trait association studies when markers are linked, or in ecological studies when explanatory variables are correlated. When tests are negatively correlated, or have a more complex dependence structure, Benjamini and Yekutieli (2001) showed that replacing m in the above procedure with

$$m \sum_{i=1}^{m} \frac{1}{i}$$

will provide FDR control. This modification is more conservative than the original procedure, and thus should be used only when made necessary by negative dependency among tests. Structure among variables



Tests ranked by p value	Bonferroni	Sequential Bonferroni	Benjamini and Hochberg
1	α / m	α / m	α / m
2	α / m	$\alpha / (m - 1)$	$2\alpha / m$
3	α / m	$\alpha / (m - 2)$	$3\alpha / m$
i	α / m	$\alpha / (m - i + 1)$	$i\alpha / m$
m	α / m	α	α

Fig. 1. Comparison of threshold p values with classical Bonferroni FWER control, Holm's sequential Bonferroni FWER control (Holm 1979, Rice 1989) and Benjamini and Hochberg FDR control (Benjamini and Hochberg 1995), when 50 tests are performed and FWER = FDR = 0.05. Threshold values are calculated according to the inset table ($m = 50$; $\alpha = 0.05$). Using Holm's sequential Bonferroni method, tests are performed from smallest to largest p until a p value exceeds the threshold; with the Benjamini and Hochberg (1995) FDR method testing is from largest to smallest p until a p value falls below the threshold.

(along a chromosome, in an environment) will most often result in positive dependencies, and thus the original correction is usually appropriate.

An example

Consider the following situation: 50 independent tests are performed, of which 15 represent true alternative hypotheses and 35 represent true null hypotheses. How is interpretation of the p values affected by using Benjamini and Hochberg (1995) FDR control instead of FWER controlling procedures? Fig. 2 shows a simulated example. P values for the true null cases were obtained by drawing randomly from a uniform distribution with boundaries 0 and 1. P values for the true alternative cases were obtained by drawing randomly from a normal distribution with a mean of 1.5 and a standard deviation of 1, and calculating the probability of drawing a more extreme value under the z distribution (thus simulating an effect size of 1.5 sd units). Note that this example serves only to illustrate application of the procedure; simulation-based power estimates are presented in Benjamini and Hochberg (1995) and Brown and Russell (1997).

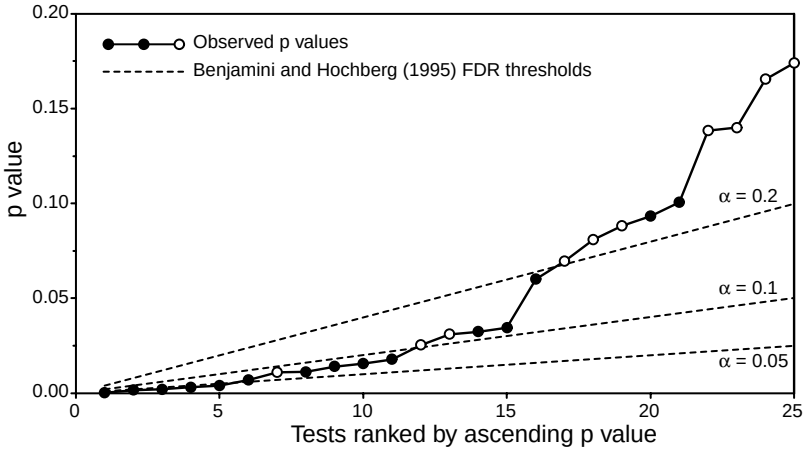
In this example, FDR control resulted in considerably fewer type II errors than either procedure for FWER control, while the number of type I errors among the significant results was close to the expected values: zero out of five at FDR 0.05 (expected value 0.25); one out of 11 at FDR 0.1 (expected value 1.1); and three out of 16 at FDR 0.2 (expected value 3.2). FWER control procedures resulted in zero type I errors at all significance levels (as expected), but at the cost of recognizing only a small fraction of the true alternative cases. A simple spreadsheet program to perform this simulation with test numbers, effect sizes and significance thresholds of choice can be downloaded as an Appendix on Oikos homepage www.oikos.lu.se/ or can be obtained from the authors.

Beyond the Benjamini and Hochberg FDR procedure

FDR control is an active field of research. Here we discuss some recent developments that either provide increased power over the 1995 Benjamini and Hochberg procedure, or provide tools for better understanding and interpretation of FDR control.

Sharpened FDR control

In common with the Holm's step-down (1979) and Hochberg's step-up (1988) sequential Bonferroni procedures, the Benjamini and Hochberg FDR method is conservative in the sense that it controls FDR no matter how many of the m tests are true null cases (m_0). The procedure, in fact, controls FDR at the level $\alpha m_0/m$ (Benjamini and Hochberg 1995). For instance, if we set $\alpha = 0.05$ and 20% of the tests happen to be true alternative cases then FDR is really controlled at a level of 0.04. The resulting loss of power can be remedied in the same spirit as sharpened FWER methods: by estimating the proportion of true null cases (m_0/m) and adjusting the critical threshold accordingly (Hochberg and Benjamini 1990). There are several graphical m_0 estimation procedures that are based on the fact that the p values of the true null cases should follow a uniform distribution while those of the true alternative cases do not. Benjamini and Hochberg (2000) present a simple sharpening procedure for their FDR method (included in our downloadable spreadsheet program),



	Truth	Decision NS	Decision S	Decision NS	Decision S	Decision NS	Decision S
		0.05		0.1		0.2	
FDR: Benjamini and Hochberg (1995)	H_0	35	*0*	34	*1*	32	*3*
	H_1	*10*	5	*5*	10	*2*	13
FDR: Sharpened Benjamini and Hochberg (2000)	H_0	34	*1*	32	*3*	29	*6*
	H_1	*8*	7	*3*	12	*0*	15
FWER: classical Bonferroni	H_0	35	*0*	35	*0*	35	*0*
	H_1	*14*	1	*12*	3	*11*	4
FWER: Holm's sequential Bonferroni	H_0	35	*0*	35	*0*	35	*0*
	H_1	*14*	1	*12*	3	*10*	5

Fig. 2. Application of Benjamini and Hochberg (1995) FDR control. The graph shows ranked p values from a simulated example with 50 tests of which 15 were true alternative hypotheses and 35 were true null hypotheses (see text); plus significance thresholds corresponding to FDR levels of 0.05, 0.1, and 0.2. Only the 25 lowest p values are shown. Closed symbols represent true alternative cases and open symbols represent true null cases. Lower panels show tabulated outcomes when applying different type I error control procedures (1995 Benjamini and Hochberg FDR control, sharpened Benjamini and Hochberg FDR control [see 'Beyond the Benjamini and Hochberg FDR procedure'], classical Bonferroni FWER control, and Holm's sequential Bonferroni FWER control), at different levels, to the simulated set of p values (H_0: true null case; H_1: true alternative case; NS: not significant; S: significant). Type I and type II errors are shown in italics.

and show that power can be increased considerably. Applied to our example of 50 tests described above, at $\alpha = 0.05$, this sharpening procedure resulted in eight type II errors and one type I error, compared to ten and zero without sharpening (Fig. 2).

In the same example, similar sharpening of the FWER controlling methods using the procedures described in Hochberg and Benjamini (1990) did not result in fewer type II errors. Brown and Russell (1997) use simulation studies to compare the power of these sharpened FWER procedures with a number of other type I error control methods, and provide software for applying the procedures to a list of p values.

The false non discovery rate (FNR)

A natural companion to the FDR is the false non discovery rate (FNR), or the expected proportion of non rejections that are incorrect (Genovese and Wasserman 2002). This is the ratio T/m − r in Table 1. The FNR plays a similar role in FDR control as power does in FWER control: given a procedure to decide which p values are significant and which are not, it quantifies a rate of making type II errors. The FNR can be estimated based on an estimate of the proportion of true null cases (m_0/m in Table 1) combined with an expectation for the number of false discoveries (from an FDR controlling procedure; V/r in Table 1, Genovese and Wasserman 2002, Taylor et al. in press). Insight into the FNR complements FDR control in a fundamental way, since the motivation to switch from FWER to FDR control is to strike a more balanced compromise between type I and type II errors. Joint consideration of FDR and FNR allows the total misclassification risk to be estimated (type I plus type II errors; Genovese and Wasserman 2002). It can be used as an evaluation tool to get a sense of the 'miss rate'. For instance, in simulation studies different FDR controlling procedures (or even different tests) can be compared in terms of their FNR. In a genomics context where thousands of tests are performed, Taylor et al. (in press) propose to estimate the FNR over a range of null hypotheses that were close to being rejected (for instance with p values between the cutoff value determined by the FDR procedure and 0.05).

Measuring significance in the FDR context

The Benjamini and Hochberg procedure sets the FDR at a desired level α, which defines a threshold to which individual p values are compared, and results in a list of rejected hypotheses of which a proportion α are expected to be type I errors. The threshold in itself does not give insight into the degree of significance for individual tests. The p value is a measure of significance in terms of the false positive rate, and is useful in the FWER context to assess for each test the risk that the null hypothesis is falsely rejected. Storey (2002) proposes a corresponding significance measure for the FDR context, the q value. This value gives the expected proportion of significant results that are truly null cases (false discoveries) when the cutoff point for H_0 rejection is at that test's p value. The q value for a test is estimated by reversing the Benjamini and Hochberg process: a rejection threshold is set at a the test's p value and the associated FDR is estimated. Q values can be calculated for each test, ranked in ascending order, and the FDR consequences of choosing a cutoff point for H_0 rejection are then apparent. Storey and Tibshirani (2003) provide software for transferring a list of p values to q values. Their FDR procedure exploits estimation of the proportion of true null hypotheses among all tests, and is more powerful than the 1995 Benjamini and Hochberg procedure and equally powerful to the sharpened 2000 Benjamini and Hochberg procedure (Black 2004).

Conclusion

When many tests are performed, keeping the proportion of false discoveries relative to all significant results at a low level is a powerful alternative to the traditional approach of avoiding even a single false discovery. Control of the FWER at α, via (sequential) Bonferroni procedures, is a suitable approach only if the penalty of making even one type I error is severe. In many studies avoiding any type I error irrespective of its cost in terms of type II errors is not a satisfactory approach. FDR control provides a sensible solution: it offers an easily interpretable mechanism to control type I errors while simultaneously allowing type II errors to be reduced.

Control of the false discovery rate is being widely adopted in genomic research. Here, genomewide scans necessitate the interpretation of hundreds or thousands of simultaneous tests, and minimizing the chance of making even a single type I error can keep the vast majority of true effects from being detected. FDR control can address a much wider range of multiple testing problems in evolution and ecology as well (García 2003, 2004), where the loss of power inherent to strict FWER control does not do justice to the nature of many experiments. FDR control is more powerful and often is more relevant than controlling the FWER. It is also flexible, and ease of interpretation is not affected by changing the significance threshold. The threshold level can vary with, for instance, the number of tests and the nature of the study (e.g. exploratory or confirmatory), in a way that is less constrained than FWER control. Sensible biological interpretation of multiple testing results may therefore benefit more from FDR than FWER control.

Acknowledgements – This material is based upon work supported by the National Science Foundation under Grant No. 9904704. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate – a practical and powerful approach to multiple testing. – J. R. Stat. Soc. B 57: 289–300.

Benjamini, Y. and Hochberg, Y. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. – J. Educ. Behav. Statist. 25: 60–83.

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

Black, M. A. 2004. A note on the adaptive control of false discovery rates. – J. R. Stat. Soc. B Met. 66: 297–304.

Brown, B. W. and Russell, K. 1997. Methods correcting for multiple testing: operating characteristics. – Stat. Med. 16: 2511–2528.

Cheverud, J. M. 2001. A simple correction for multiple comparisons in interval mapping genome scans. – Heredity 87: 52–58.

García, L. V. 2003. Controlling the false discovery rate in ecological research. – Trends Ecol. Evol. 18: 553–554.

García, L. V. 2004. Escaping the Bonferroni iron claw in ecological studies. – Oikos 105: 657–663.

Genovese, C. and Wasserman, L. 2002. Operating characteristics and extensions of the false discovery rate procedure. – J. R. Stat. Soc. B 64: 499–517.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. – Biometrika 75: 800–802.

Hochberg, Y. and Benjamini, Y. 1990. More powerful procedures for multiple significance testing. – Stat. Med. 9: 811–818.

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

Moran, M. D. 2003. Arguments for rejecting the sequential Bonferroni in ecological studies. – Oikos 100: 403–405.

Rice, W. R. 1989. Analyzing tables of statistical tests. – Evolution 43: 223–225.

Schweder, T. and Spjøtvoll, E. 1982. Plots of P-values to evaluate many tests simultaneously. – Biometrika 69: 493–502.

Storey, J. D. 2002. A direct approach to false discovery rates. – J. R. Stat. Soc. B 64: 479–498.

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

Taylor, J., Tibshirani, R. and Efron, B. in press. The 'miss rate' for the analysis of gene expression data.–Biostatistics.