

1 **Implementing false discovery rate control: increasing your power**

2

3 Koen J.F. Verhoeven¹, Katy L. Simonsen² and Lauren M. McIntyre¹

4

5 ¹ Computational Genomics, Department of Agronomy, Purdue University, Lilly Hall of

6 Life Sciences, 915 W. State Street, West Lafayette, IN 47907-2054. Fax: +1 765 496

7 2926

8

9 ² Department of Statistics, Purdue University, 150 N. University Ave. West Lafayette, IN

10 47907-2068

11

12

13 Corresponding author: K.J.F. Verhoeven, email: kverhoeven@purdue.edu

1 **Abstract**

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

1 **The problem**

2 The appropriate threshold to declare a test statistic's p value significant becomes
3 complex when more than one test is performed. In the absence of a true effect each test
4 has a chance of α to yield a significant result, and the chance of drawing at least one false
5 conclusion increases rapidly with the number of tests performed. Protection against false
6 rejections of the null hypothesis, or type I errors, is usually achieved via Bonferroni-type
7 correction (e.g. Holm 1979). By performing individual tests at error rates that are a
8 fraction of the overall nominal α , the chance of making even a single type I error can be
9 maintained at the desired α level (usually 5%). This is called control of the familywise
10 error rate (FWER). With an increasing number of tests, maintaining a low chance of
11 making even one type I error comes at the direct cost of making more type II errors, i.e.
12 not recognizing a true effect as significant. The classical Bonferroni procedure, which
13 performs each of m tests at a type I error rate of α/m , is undesirable because of this trade-
14 off: only a very strong effect is likely to be recognized as significant when many tests are
15 performed. Several improvements to the classical Bonferroni have been proposed in order
16 to reduce the problem of low power (see Garcia 2004 for a recent overview). For
17 instance, the well-known Holm's step-down or sequential Bonferroni procedure (Holm
18 1979, popularized among evolutionary biologists and ecologists by Rice 1989) performs
19 tests in order of increasing p values, and conditional on having rejected tests with smaller
20 p values an increasingly permissive threshold can be used while maintaining the FWER
21 at the desired level (5%). Further power gains are possible with the sequential approach
22 by using a step-up instead of a step-down procedure (that is, testing in order of decreasing
23 p values, Hochberg 1988), by estimating the number of true null hypotheses and

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

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

16

17 **Controlling the false discovery rate**

18 An elegant way to deal with the problem, that was recently advocated for ecological
19 studies by García (2003, 2004), is to control the proportion of significant results that are
20 in fact type I errors ('false discoveries') instead of controlling the chance of making even
21 a single type I error. This new approach was developed by Benjamini and Hochberg
22 (1995). To see the difference between FWER and the false discovery rate (FDR),
23 consider the potential outcomes of each test (Table 1). FWER is the probability that V ,

1 the number of type I errors, is greater than or equal to one. FDR, as defined by Benjamini
2 and Hochberg (1995), is the expected proportion of type I errors among all significant
3 results (V/r). Control of FWER, for instance via Bonferroni or sequential Bonferroni
4 adjustment of the per comparison error rate, means that the probability that $V \geq 1$ is
5 maintained at a desired level. Control of FDR means that the expected proportion V/r is
6 maintained at a desired level. When all null hypotheses are true, controlling FWER and
7 FDR are equivalent. In that case either $V/r = 0$ (by definition if $V = 0$) or $V/r = 1$ (if $V > 0$,
8 because all significant results are false), and the *expected* ratio equals the chance that any
9 false rejection is made. However, if some of the alternative hypotheses are true and $S > 0$,
10 then V/r is either 0 (if $V = 0$) or $0 < V/r < 1$ (if $V > 0$), and the expected ratio is smaller
11 than the chance that any false rejection is made (see Benjamini and Hochberg 1995). In
12 those cases FDR is smaller than FWER, and controlling FDR at, say, 5% can result in
13 fewer type II errors than controlling FWER at 5%. The gain increases when more
14 alternative hypotheses are true.

15

16 The following simple procedure to control FDR at level α was proposed by Benjamini
17 and Hochberg (1995): For m tests, rank the p values in ascending order $P_{(1)} \leq P_{(2)} \leq \dots \leq$
18 $P_{(m)}$, and denote by $H_{(i)}$ the null hypothesis corresponding to $P_{(i)}$; Let k be the largest i for
19 which

20

21

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

22

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

10 The above procedure was shown by Benjamini and Hochberg (1995) to control FDR
11 in the case when all m tests are independent, and was also shown (Benjamini and
12 Yekutieli 2001) to control FDR when tests are positively correlated. These are thought to
13 be the most common situations in genetic and ecological studies. Positive dependence,
14 for instance, can occur in marker-trait association studies when markers are linked, or in
15 ecological studies when explanatory variables are correlated. When tests are negatively
16 correlated, or have a more complex dependence structure, Benjamini and Yekutieli
17 (2001) showed that replacing m in the above procedure with

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

18
19
20
21 will provide FDR control. This modification is more conservative than the original
22 procedure, and thus should be used only when made necessary by negative dependency
23 among tests. Structure among variables (along a chromosome, in an environment) will

1 most often result in positive dependencies, and thus the original correction is usually
2 appropriate.

3

4

5 **An example**

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

17 In this example, FDR control resulted in considerably fewer type II errors than either
18 procedure for FWER control, while the number of type I errors among the significant
19 results was close to the expected values: zero out of five at FDR 0.05 (expected value
20 0.25); one out of 11 at FDR 0.1 (expected value 1.1); and three out of 16 at FDR 0.2
21 (expected value 3.2). FWER control procedures resulted in zero type I errors at all
22 significance levels (as expected), but at the cost of recognizing only a small fraction of
23 the true alternative cases. A simple spreadsheet program to perform this simulation with

1 test numbers, effect sizes and significance thresholds of choice can be downloaded at
2 <http://www.genomics.purdue.edu/people/koen/>, or can be obtained from the authors.

3

4 **Beyond the Benjamini and Hochberg FDR procedure**

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

8

9 *Sharpened FDR control*

10 In common with the Holm's step-down (1979) and Hochberg's step-up (1988)
11 sequential Bonferroni procedures, the Benjamini and Hochberg FDR method is
12 conservative in the sense that it controls FDR no matter how many of the m tests are true
13 null cases (m_0). The procedure, in fact, controls FDR at the level $\alpha m_0/m$ (Benjamini and
14 Hochberg 1995). For instance, if we set $\alpha=0.05$ and 20% of the tests happen to be true
15 alternative cases then FDR is really controlled at a level of 0.04. The resulting loss of
16 power can be remedied in the same spirit as sharpened FWER methods: by estimating the
17 proportion of true null cases (m_0/m) and adjusting the critical threshold accordingly
18 (Hochberg and Benjamini 1990). There are several graphical m_0 estimation procedures
19 that are based on the fact that the p values of the true null cases should follow a uniform
20 distribution while those of the true alternative cases do not. Benjamini and Hochberg
21 (2000) present a simple sharpening procedure for their FDR method (included in our
22 downloadable spreadsheet program), and show that power can be increased considerably.
23 Applied to our example of 50 tests described above, at $\alpha = 0.05$, this sharpening

1 procedure resulted in eight type II errors and one type I error, compared to ten and zero
2 without sharpening (Fig. 2).

3 In the same example, similar sharpening of the FWER controlling methods using the
4 procedures described in Hochberg and Benjamini (1990) did not result in fewer type II
5 errors. Brown and Russell (1997) provide software for applying these sharpened FWER
6 (and other) correction procedures to a list of p values.

7 8 *The false non-discovery rate (FNR)*

9 A natural companion to the FDR is the false non-discovery rate (FNR), or the
10 expected proportion of non-rejections that are incorrect (Genovese and Wasserman
11 2002). This is the ratio $T/m-r$ in Table 1. The FNR plays a similar role in FDR control as
12 power does in FWER control: given a procedure to decide which p values are significant
13 and which are not, it quantifies a rate of making type II errors. The FNR can be estimated
14 based on an estimate of the proportion of true null cases (m_0/m in Table 1) combined with
15 an expectation for the number of false discoveries (from an FDR controlling procedure;
16 V/r in Table 1) (Genovese and Wasserman 2002, Taylor et al. in press). Insight into the
17 FNR complements FDR control in a fundamental way, since the motivation to switch
18 from FWER to FDR control is to strike a more balanced compromise between type I and
19 type II errors. Joint consideration of FDR and FNR allows the total misclassification risk
20 to be estimated (type I plus type II errors; Genovese and Wasserman 2002). It can be
21 used as an evaluation tool to get a sense of the ‘miss rate’. For instance, in simulation
22 studies different FDR controlling procedures (or even different tests) can be compared in
23 terms of their FNR. In a genomics context where thousands of tests are performed, Taylor

1 et al. (in press) propose to estimate the FNR over a range of null hypotheses that were
2 close to being rejected (for instance with p values between the cutoff value determined by
3 the FDR procedure and 0.05).

4

5 *Measuring significance in the FDR context*

6 The Benjamini and Hochberg procedure sets the FDR at a desired level α , which
7 defines a threshold to which individual p values are compared, and results in a list of
8 rejected hypotheses of which a proportion α are expected to be type I errors. The
9 threshold in itself does not give insight into the degree of significance for individual tests.
10 The p value is a measure of significance in terms of the false positive rate, and is useful
11 in the FWER context to assess for each test the risk that the null hypothesis is falsely
12 rejected. Storey (2002) proposes a corresponding significance measure for the FDR
13 context, the q value. This value gives the expected proportion of significant results that
14 are truly null cases (false discoveries) when the cutoff point for H_0 rejection is at that
15 test's p value. The q value for a test is estimated by reversing the Benjamini and
16 Hochberg process: a rejection threshold is set at a the test's p value and the associated
17 FDR is estimated. Q values can be calculated for each test, ranked in ascending order,
18 and the FDR consequences of choosing a cutoff point for H_0 rejection are then apparent.
19 Story and Tibshirani (2003) provide software for transferring a list of p values to q
20 values. Their FDR procedure exploits estimation of the proportion of true null hypotheses
21 among all tests, and is more powerful than the 1995 Benjamini and Hochberg procedure
22 and equally powerful to the sharpened 2000 Benjamini and Hochberg procedure (Black
23 2004).

1

2

3 **Conclusion**

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

12 Control of the false discovery rate is being widely adopted in genomic research. Here,
13 genomewide scans necessitate the interpretation of hundreds or thousands of
14 simultaneous tests, and minimizing the chance of making even a single type I error can
15 keep the vast majority of true effects from being detected. FDR control can address a
16 much wider range of multiple testing problems in evolution and ecology as well (Garcia
17 2003, 2004), where the loss of power inherent to strict FWER control does not do justice
18 to the nature of many experiments. FDR control is more powerful and often is more
19 relevant than controlling the FWER. It is also flexible, and ease of interpretation is not
20 affected by changing the significance threshold. The threshold level can vary with, for
21 instance, the number of tests and the nature of the study (e.g. exploratory or
22 confirmatory), in a way that is less constrained than FWER control. Sensible biological

1 interpretation of multiple testing results may therefore benefit more from FDR than
2 FWER control.

3

4 **Acknowledgements**

5 This work was supported by the National Science Foundation, grant NSF DBI-
6 9904704.

7

8 **References**

9

- 10 Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate - a practical
11 and powerful approach to multiple testing. - J. Roy. Stat. Soc. B 57: 289-300.
- 12 Benjamini, Y. and Hochberg, Y. 2000. On the adaptive control of the false discovery rate
13 in multiple testing with independent statistics. - J. Educ. Behav. Statist. 25: 60-83.
- 14 Benjamini, Y. and Yekutieli, D. 2001. The control of the false discovery rate in multiple
15 testing under dependency. - Ann. Stat. 29: 1165-1188.
- 16 Black, M. A. 2004. A note on the adaptive control of false discovery rates. - J. Roy. Stat.
17 Soc. B Met. 66: 297-304.
- 18 Brown, B. W. and Russell, K. 1997. Methods correcting for multiple testing: operating
19 characteristics. - Stat. Med. 16: 2511-2528.
- 20 Cheverud, J. M. 2001. A simple correction for multiple comparisons in interval mapping
21 genome scans. - Heredity 87: 52-58.
- 22 Garcia, L. V. 2003. Controlling the false discovery rate in ecological research. - Trends
23 Ecol. Evol. 18: 553-554.

1 Garcia, L. V. 2004. Escaping the Bonferroni iron claw in ecological studies. - *Oikos* 105:
2 657-663.

3 Genovese, C. and Wasserman, L. 2002. Operating characteristics and extensions of the
4 false discovery rate procedure. - *J. Roy. Stat. Soc. B* 64: 499-517.

5 Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. -
6 *Biometrika* 75: 800-802.

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

9 Holm, S. 1979. A simple sequentially rejective multiple test procedure. - *Scand. J. Stat.* 6:
10 65-70.

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

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

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

16 Storey, J. D. 2002. A direct approach to false discovery rates. - *J. Roy. Stat. Soc. B* 64:
17 479-498.

18 Storey, J. D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. -
19 *Proc. Natl. Acad. Sci. U.S.A.* 100: 9440-9445.

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

22

1 **Figure legends**

2

3 Figure 1. Comparison of threshold p values with classical Bonferroni FWER control,
4 Holm's sequential Bonferroni FWER control (Holm 1979, Rice 1989) and Benjamini &
5 Hochberg FDR control (Benjamini and Hochberg 1995), when 50 tests are performed and
6 FWER = FDR = 0.05. Threshold values are calculated according to the inset table ($m =$
7 50; $\alpha = 0.05$). Using Holm's sequential Bonferroni method, tests are performed from
8 smallest to largest p until a p value exceeds the threshold; with the Benjamini &
9 Hochberg (1995) FDR method testing is from largest to smallest p until a p value falls
10 below the threshold.

11

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

1

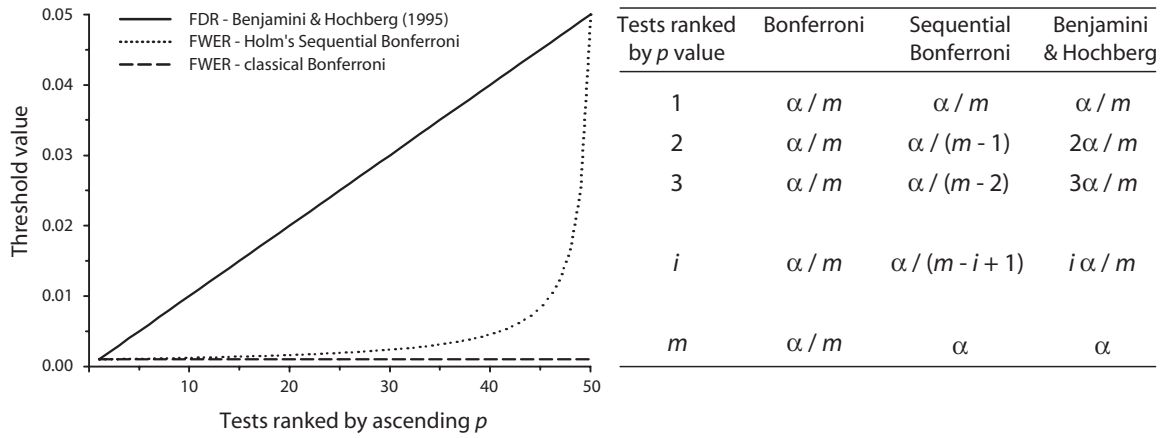
2 Table 1. Possible outcomes of individual tests. Note that V is the number
3 of type I errors; T is the number of type II errors; and only m , r and $m-r$
4 are observed while the other variables are unknown.

5

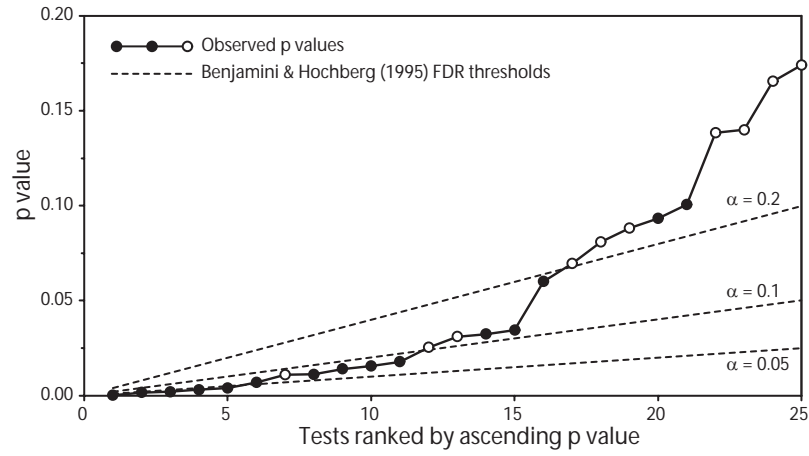
Truth	Decision		Total
	Not significant	Significant	
Null hypothesis	U	V	m_0
Alternative hypothesis	T	S	$m-m_0$
Total	$m-r$	r	m

6

1 | Figure 1.



1 Figure 2



		Decision					
		0.05		0.1		0.2	
FDR: Benjamini & Hochberg (1995)	H ₀	35	0	34	1	32	3
	H ₁	10	5	5	10	2	13
FDR: Sharpened Benjamini & 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