ELSEVIER

Behavioural Brain Research 125 (2001) 279–284

BEHAVIOURAL BRAIN RESEARCH

www.elsevier.com/locate/bbr

Controlling the false discovery rate in behavior genetics research

Yoav Benjamini [a,*], Dan Drai [b], Greg Elmer [c], Neri Kafkafi [d], Ilan Golani [b]

[a] *Department of Statistics and O.R., The Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel*
[b] *Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel*
[c] *Maryland Psychiatry Research Institute, Baltimore, MD, USA*
[d] *National Institute of Drug Abuse, Baltimore, MD, USA*

Received 5 July 2000; accepted 16 March 2001

Abstract

The screening of many endpoints when comparing groups from different strains, searching for some statistically significant difference, raises the multiple comparisons problem in its most severe form. Using the 0.05 level to decide which of the many endpoints' differences are statistically significant, the probability of finding a difference to be significant even though it is not real increases far beyond 0.05. The traditional approach to this problem has been to control the probability of making even one such error—the Bonferroni procedure being the most familiar procedure achieving such control. However, the incurred loss of power stemming from such control led many practitioners to neglect multiplicity control altogether. The False Discovery Rate (FDR), suggested by Benjamini and Hochberg [J Royal Stat Soc Ser B 57 (1995) 289], is a new, different, and compromising point of view regarding the error in multiple comparisons. The FDR is the expected proportion of false discoveries among the discoveries, and controlling the FDR goes a long way towards controlling the increased error from multiplicity while losing less in the ability to discover real differences. In this paper we demonstrate the problem in two studies: the study of exploratory behavior [Behav Brain Res (2001)], and the study of the interaction of strain differences with laboratory environment [Science 284 (1999) 1670]. We explain the FDR criterion, and present two simple procedures that control the FDR. We demonstrate their increased power when used in the above two studies. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Multiple comparisons; Exploratory behavior in mice; Bonferroni procedure; FDR

1. Introduction

A quantifiable description of mouse behavior should promote the mapping of the mouse genome by characterizing the repertoires of inbred strains, congenic lines, knockouts, transgenic lines, and populations obtained by selective breeding. The need for such characterization has resulted in the design of batteries of behavioral and physiological tests. Such studies never constitute of a single pre-specified measure, which is being compared between two strains of mice. The studies develop and explore many characteristics—also called behavioral endpoints, trying to identify those endpoints for which there is a significant strain difference. We estimate that at the time of the writing of this paper the working list of behavioral endpoints is about a 100 endpoints long, and keeps growing.

The screening of many endpoints when comparing groups from different strains, searching for some statistically significant difference, raises the multiple comparisons problem in its most severe form. The search is conducted by testing each hypothesis of no strain difference in some endpoint, which is done at some declared level of statistical significance, say at the 0.05. Detecting such a difference as 'statistically significant' amounts to making a statistical discovery. But then, when screening such a large family of hypotheses simultaneously, the probability of making a false discovery may increase far beyond the declared 0.05 level. If 100 endpoints are compared in a study, assuming there are few real trait differences between the strains, and if no action is taken, the average number of errors per study will be a little less than 100×0.05, i.e. close to 5. This

* Corresponding author. Tel.: +972-3-6408756; fax: +972-3-6409357.
E-mail address: ybenja@post.tau.ac.il (Y. Benjamini).

0166-4328/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.
PII: S0166-4328(01)00297-2

will be the case whether the endpoints are statistically independent or not.

The traditional approach in multiple hypotheses testing to tackle this increased probability of making false discoveries, has been to control the probability of making even one false discovery—the control of the *familywise error-rate* as it is called in the statistical jargon. The books by Hochberg and Tamhane [9], Westfall and Young [15] and Hsu [10] all reflect this tradition. The control of this error-rate at some level α requires each of the m tests of the endpoints to be conducted at lower levels. In the Bonferroni procedure, for example, α/m has to be used.

The Bonferroni procedure is just an example, as more powerful procedures that control the probability of making even one false discovery are currently available for many multiple comparison problems. Many of the newer procedures are as flexible as the Bonferroni, making use of the P-values only. For a recent review see [9]. Still, there is a fundamental drawback to this traditional approach: the probability to discover a real strain difference in an endpoint (the *power*) is greatly reduced when screening a large family of potential endpoints. The incurred loss of power in large problems (even with the newer procedures) led many practitioners to neglect multiplicity control altogether. While mandatory in psychological research, most medical journals do not require the analysis of the multiplicity effect on the statistical conclusions, the leading New England Journal of Medicine being among the other few.

In genetic research, the need for multiplicity control has been debated heavily. In QTL analysis, the debate resulted in some compromise. Allow the probability of making even one false discovery to be as high as half in order to increase power, then follow the original study with a more limited confirmatory study to ensure better protection against false discoveries (see [14] for background and further references). This strategy is elsewhere advocated in order to deal with the results of multiplicity in smaller studies, and can be quite an effective one. Nevertheless it has shortcomings: it is usually not possible to quantify the properties of the discoveries made in the follow-up study; and it turns out to be wasteful if no multiplicity adjustment is offered at the first stage. Another unfortunate practical problem is that occasionally the second stage is not performed at all. In very large and costly studies all three problems tend to appear.

It should be emphasized that the recent trend away from hypotheses testing towards confidence statements does not solve the multiplicity problem. In most analyses a decision about the statistical significance is reached by looking whether zero difference is included in the confidence interval or not—taking us back to the same multiplicity problem.

The False Discovery Rate (FDR) is a new and different point of view at how the errors in multiple comparisons could be considered [3]. The FDR is the expected proportion of false discoveries among the discoveries. In this paper we shall explain this notion and discuss some simple procedures that control the FDR. We stress the importance of controlling for the treacherous effect of multiplicity, while not being overly conservative.

2. Two motivating example

In a separate paper in this issue, Drai et al. [8] propose to study the open field behavior of mice using the approach developed in the study of rats. They describe an effort to augment the commonly used measures of the open field test with a set of new ethologically relevant parameters. These parameters, which can be measured automatically and efficiently, reveal a natural structure that involves motivation, navigation, spatial memory and learning. Some 17 such parameters are identified in that study, and are presented in the leftmost column of Table 1. The values of these parameters were estimated and compared between eight male C57BL/6Jtau (C57) Bulb, and eight male BALB/cJtau (BALB) mice from the Tel Aviv University medical school stocks. We use those results to motivate the approach and demonstrate the procedures involved. See [8] for more detailed description of the experimental setting.

Ignoring the issue of multiplicity altogether, all 10 hypotheses for which the observed p-value is less than 0.05 should be rejected. Using the Bonferroni procedure, each p-value is compared to $0.05/17 = 0.0029$. In this case only six differences are statistically significant. There is quite a difference in this study between the implications of the two approaches.

A somewhat similar situation appears in the work of Crabbe et al. [7], who studied the possible confounding influences of laboratory environment on tests of mice behavior. Some 56 statistical hypotheses are tested, eight among them being the interaction between strain differences and laboratories effects. Of these 56 only 14 would be determined as statistically significant at the 0.05 level if a multiple comparison adjustment would have been taken using the Boferroni procedure. The authors argue in their published paper that p-values should not be adjusted for multiplicity, and their Web site discusses the rationale for their position.

Nevertheless, they do take a partial step in face of the multiplicity. Instead of the conventional 0.05 level used for statistical significance, they chose a somewhat stricter level, namely the 0.01 level.

Table 1
The results of comparing 17 exploratory behavior measures between eight C57 and eighht BALB mice

Measure	Observed P-values	Rank (i)	Bonferroni threshold	FDR (BH) thresholds	FDR (BL) thresholds
Lingering time (prop.)	0.000001	1	0.0029	0.0029	0.0029 Start here
Lingering speed (cm/s)	0.000013	2	0.0029	0.0058	0.0033
Early activity in move segments (m)	0.000065	3	0.0029	0.0088	0.0037
Early activity (m)	0.00063	4	0.0029	0.0117	0.0043
Spread of lingering (cm)	0.0008	5	0.0029	0.0147	0.0050
Dynamics of activity	0.0017	6	**0.0029**	0.0176	0.0059
Dynamics of diversity	0.0032	7	0.0029	0.0205	0.0070
Number of excursions	0.0065	8	0.0029	0.0235	**0.0085**
Movement speed (cm/s)	0.0148	9	0.0029	**0.0264**	0.0104
Spread of move segments	0.049	10	0.0029	0.0294	0.0132
Stops per excursions (upper quartile)	0.094	11	0.0029	0.0323	0.0173
Center activity (prop.)	0.11	12	0.0029	0.0352	0.0236
Center rest (prop.)	0.15	13	0.0029	0.0382	0.0340
Activity (m)	0.24	14	0.0029	0.0411	0.05
Lingering activity (prop.)	0.45	15	0.0029	0.0441	0.05
Diversity	0.56	16	0.0029	0.047	0.05
Lingering at home base (prop.)	0.87	17	0.0029	0.05 Start here	0.05

The list of the observed P-value (t-test after appropriate transformation or Wilcoxon test) is sorted from smallest to largest. The rightmost three columns demonstrate three different multiple comparisons procedures: the Bonferroni procedure, the two FDR controlling procedures of Benjamini and Hochberg (BH) and of Benjamini and Liu (BL).

3. The false discovery rate criterion (FDR)

Consider the case of m endpoints being compared between two strains, or more generally any family of m null hypotheses being tested in a study. Some tested null hypotheses of no difference may be true—possibly even all—meaning no difference exists between the two strains in the corresponding endpoints. Other hypotheses of no difference may be false—meaning real differences exist—and we wish to discover these real differences as *statistically significant*, granting us with statistical discoveries. Obviously we would like to discover as many as possible of the real differences as such, while making as few as possible errors of falsely discovering a difference which is not real.

However, in a statistical study, it is quite unavoidable that when we find a number of differences as statistically significant, some unknown number of false discoveries will creep in. Let us 'measure the harm' imposed by such errors, by considering the proportion of the false discoveries among the discoveries. In the case that no discovery was made this proportion is defined as 0, since there is no way for any harm to arise from false discoveries. It makes sense to control this proportion at some desired level in each and every study, but this is impossible. Consider instead the average value of this proportion, which we define as the *False Discovery Rate* (*FDR*). This false discovery rate can be controlled at any desired level.

The FDR criterion is a compromise between the unadjusted analysis of the multiple tests, and the traditionally adjusted approaches. If there is no behavioral difference between the strains whatsoever, that is all tested hypotheses are true, controlling the FDR controls the traditional probability of making even one false discovery. Therefore, it also makes sense to use the conventional levels such as 0.05 or 0.01 for FDR control (though in some applications higher values may be justifiable). However, when many of the endpoints are discovered to be different, indicating that many differences are real, the error from a single false discovery is not always as crucial for drawing conclusions from the entire study, and the proportion of errors among the discoveries is controlled instead. Thus we are ready to bear with more errors when many discoveries are made, but with fewer errors when fewer discoveries are made: two error out of 40 established differences is bearable, two errors out of four is certainly not.

In many applied problems it has been argued that the control of the FDR at some specified level is the more appropriate response to the multiplicity concern: in educational research [16], signal processing [1], Medical Research [12], in Psychology [11] and in Genetics [14]. The practical difference between the two approaches is not small or trivial, and the larger the problem the more dramatic the difference is. In the following section we present two procedures that control the FDR at the desired level.

4. Two FDR controlling procedures

Benjamini and Hochberg provide in [2] a simple stepwise procedure (BH) that controls the FDR when the test statistics are statistically independent. This procedure has been lately shown to control the FDR when the test statistics are positively correlated as well. The procedure makes use of the observed significance level (the p-values) only. It is available in SAS (where it is called the FDR procedure), but once the P-value are available from any statistical software, the extra calculation can be done easily within a spreadsheet software such as excel using the built-in functions, or even can be performed by hand, as we show below. We shall also present the procedure of Benjamini and Liu [5] (BL), that is a modification of [4] that always controls the FDR—even for generally correlated test statistics.

As mentioned above, both procedures make use of the p-values of the tested differences only, so the statistical test itself may be tailored to the problem at hand, be it t-test, binomial test, χ^2-test, or some other non-parametric test. The individual P-values should then be sorted from smallest to largest as is demonstrated in columns 2 and 3 of Table 1. Denote the i-th smallest P-value (in the i-th row) by $p_{(I)}$, for each i between 1 and m. The BH procedure in [2] runs as following:

Starting from the largest P-value $P_{(m)}$, compare $P_{(m)}$ with $0.05 \times i/m$. Continue as long as $P_{(I)} > 0.05 \times i/m$. Let k be the first time when $P_{(k)}$ is less than or equal to $0.05 \times k/m$, and declare the differences corresponding to the smallest k P-values as significant.

The procedure for controlling the FDR at level 0.05 is demonstrated by calculating the relevant constants and showing them in column 5 of Table 1. Starting with the largest P-value 0.87, which is in row 17 since it is the 17th in order, compare it with $0.05 \times 17/17$, which is simply 0.05. Finding that $0.87 > 0.05$ we continue our search one row up-with the 16th P-value, which is 0.56. We compare it to $0.05 \times 16/17$, which is 0.047, seen again in column 5. The P-value is still larger than the respective constant, so we continue further up in Table 1 to row 15 and so on. The first time for which the inequality is reversed is when the nineth P-value, which is 0.0148, is less than $0.05 \times 9/17$, which is 0.0264. We thus stop at the nineth row, and reject all nine differences for which the P-values is equal or less than 0.0264.

Note that in this procedure we start at the 0.05 level, and if all differences are statistically significant at this level—all are rejected, as if no adjustment for multiplicity was taken. If we have to climb all the way up to the smallest P-value, that difference will be significant only if its P-value is equal or less than 0.05/17, as if the Bonferroni procedure was used. The in between the constants are linearly spaced.

(An intuitive explanation of why these constants achieve FDR control can be given: if we reject for P-values less than $P_{(k)}$, the average number of false discoveries is $P_{(k)}\, m$ and their number is k. A crude bound for FDR is $P_{(k)}\, m/k$. For this to be less than q, $P_{(k)}$ has to be less than qk/m.)

The BL procedure in [5] runs as following:

Starting from the smallest P-value $P_{(1)}$, compare each $P_{(i)}$ with $h_{(i)} = \min(0.05, 0.05 \times m/(m+1-i)^2)$. Reject the hypothesis corresponding to $P_{(1)}$ if it is smaller than or equal to the threshold $h_{(1)}$, and continue to reject the hypotheses as long as $P_{(i)}$ is less than or equal to $h_{(i)}$. Stop when $P_{(k)} > h_{(k)}$ for the first time. Reject all of the hypotheses corresponding to the smallest $k-1$ P-values.

This procedure for controlling the FDR at level 0.05 is demonstrated by calculating the relevant constants $h_{(i)}$ and showing them in column 6 of Table 1. Starting with the smallest P-value 0.000001 which is in row 1, compare it with $\min(0.05, 0.05 \times 17/(17+1-1)^2)$, which is 0.0029. Finding that 0.000001 is less than 0.0029, we continue to the second smallest P-value, in row 2, which is 0.000013. We compare it to $0.05 \times 17/(17+1-2)^2$, which is $0.05 \times 17/(16)^2$ i.e. to 0.0033. The P-value is still smaller than the respective constant, so we continue down Table 1. The last time when the the P-value is smaller than the respective constant is when the eighth P-value which is 0.0065 is less than $0.05 \times 17/(17+1-8)^2$, which is 0.0085. For the nineth P-value $\min(0.05, 0.05 \times 17/(17+1-9)^2) = 0.0104$, and the nineth P-value is 0.0148 which is bigger. We therefore stop, and declare a real difference only in the eight endpoints for which the P-values which are equal or smaller than 0.0085.

Note that the largest four constants are all 0.05 because $0.05 \times m/(m+1-i)^2$ is larger than 0.05. Following the remark in [5], this modification of the procedure ensures that a difference is not declared as statistically significant unless: (a) the FDR is less than 0.05 for the set of declared discoveries; and (b) individually its statistical significance is at least 0.05.

Again, as with the first procedure, at the two extremes the P-values are compared to 0.05 and to 0.05/17. The progression of the thresholds, though, is not linear, and the stepping direction is in the other direction—stepping down from the first row with the smallest P-value to the largest.

So far we have demonstrated both procedures on the same data. There is a difference between the results of the analysis of the two procedures, the first one finding nine endpoints with strain difference, the second only eight. Which procedure is more appropriate? The first procedure requires that the tests be based on behavioral endpoints that are either statistically independent or positively dependent [6]. The second one requires no such assumption—in fact it requires no assumption at

all. Checking the dependencies among our differences of endpoints we found a few large and significant negative correlations. Therefore, the correct procedure to use in this example is the more general second procedure.[1]

Two endpoints were added to the list of differences after using the FDR controlling procedure: the number of excursions, and dynamics of diversity. The difference discovered in the second endpoint is of special interest, as 'diversity' reflects the spatial and temporal spread of stops within a specified time window. Diversity is low when animals show strereotyped behavior, and high when they show unrestrained free behavior. 'Dynamics of Diversity' compares this measure in the first and second halves of the session. In C57 mice diversity is high as soon as they are introduced into the arena. In contrast, in Balb mice there is a buildup of diversity across the session. The significant difference between the two strains may indicate a corresponding biologically important difference in their cognitive spatial behavior.

As to the second example from [7], the design of that study implies that the tests are almost independent (testing interactions and main effects in a balanced ANOVA). Therefore, we may use the BH procedure. Exact P-values are not given in that paper, but the available information allows one to make some rough calculations: among the 56 tested hypotheses, 14 P-values are less than 0.00001, six between 0.00001 and 0.001, and four between 0.001 and 0.01. Thus the 24th ordered P-value is less than 0.01. Since $0.01 < 0.05 \times 24/56$ at least these 24 P-values should be rejected. Had we had the full information, it is quite possible that a few more hypotheses could be rejected.

This paper also demonstrates how to overcome a possible manipulation of the FDR criterion. One may make the FDR criterion less restrictive by 'throwing in' among the tested differences a few endpoints for which the differences are already known to be real and large, and therefore the significant conclusion about them is almost certain. This implies in turn that it will also become easier than before to discover more questionable endpoints because their P-values will be now compared to larger constants.

In the Crabbe et al. study [7], the strains and behavior traits were so chosen to bring out strain differences as clearly as possible—here for a good reason and not merely in order to manipulate the FDR criterion. The result may still be the same, that the procedure has found more discoveries than appropriate. Luckily, we can study the implications of such decisions. In the Crabbe et al.'s study the eight 'obvious' strain main effects are clearly identified. We can remove them from the analysis thereby leaning towards a more conservative FDR analysis. Now there are 16 P-values less than 0.01, each of these P-values should be compared to $0.05 \times 16/48$ and these same 16 hypotheses are still rejected. The skeptic readers of other studies can always perform the above sensitivity analysis, if they suspect a problem, because the relevant information has to be included in studies that control the FDR.

5. Discussion

It is clear that the multiple comparisons problem has to be addressed in the comparison of behavioral endpoints between strains of mice. This is especially important in any automated screening tool that is designed for discovering of genetic differences, as in the study of exploratory behavior [8]. In that study 10 differences would be found significant if no multiplicity adjustment were taken, six if the traditional Bonferroni was used. Two endpoints were added to the six significant differences after using the FDR controlling procedure: the number of excursions, and dynamics of diversity. As discussed above, the strain difference discovered in the second endpoint is of special interest.

The control of the false discovery rate seems to us the appropriate approach for the purpose, striking a balance between the concern about making too many false discoveries and the concern about missing the discovery of a real difference that may arise from being too conservative. In the analysis of Crabbe et al. [7] the FDR controlling procedure actually got automatically at the compromising level of strictness chosen by the experimenters on an intuitive base. We therefore recommend using FDR controlling procedures for screening an established list of endpoints or a potential pool of new ones.

Traditional multiple comparisons procedures offer even stricter control against the increased probability of discovering a non-existing difference. Thus, if differences are found which pass the Bonferroni threshold, or other procedures that control the familywise error-rate, the evidence should be regarded as stronger. Williams, Jones and Tukey [16] suggest calling such differences as 'highly significant', and those passing the FDR threshold as simply 'significant'. We are not sure that such formalism is needed, but we do emphasize that if a difference is not found to be statistically significant after controlling for the FDR, it should not be declared 'statistically significant' at all—even if individually its corresponding P-value is less than 0.05.

Two procedures were demonstrated, one for independent and positively dependent test statistics, where the thresholds increase linearly, the other for general dependency. If the assumptions justify using the first

[1] This reflects the currently established knowledge. Some initial results suggest that the first procedure can be used even for negative correlations when the tests are two-sided, as is the case here.

procedure, and the number of tested hypotheses is moderate to large (say bigger than 8), the first procedure should be preferred. Other FDR controlling procedures have been developed, and research about newer methodologies continues. Some of the newer procedures are more powerful than the above two: Benjamini and Hochberg [3] give an adaptive procedure for independent test statistics, and Yekutieli and Benjamini [17] offer a re-sampling procedure utilizing the dependency structure of the data. Troendle [13] designs procedures for the special case where many comparisons are made with a single control. For recent developments in FDR methodology, references, and statistical software consult with http://www.math.tau.ac.il/ ~ ybenja, which is being regularly updated. Even without further consulting about other procedures, with the aid of the two procedures offered in this paper researchers in behavioral genetics could actually use the FDR approach to adjust for the multiplicity effect in their regular work.

Acknowledgements

This study is part of the project 'Phenotyping mouse exploratory behavior' supported by NIH 1 R01 NS40234-01.

References

[1] Abramovich F, Benjamini Y. Adaptive thresholding of wavelet coefficients. Comput Stat Data An 1996;22(4):351–61.
[2] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. J Royal Stat Soc Ser B 1995;57(1):289–300.
[3] Benjamini Y, Hochberg Y. The adaptive control of the false discovery rate in multiple comparison problems. J Educ Behav Stat 2000;25(1):60–83.
[4] Benjamini Y, Liu W. A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. J Stat Plann Inference 1999;82(1-2):163–70.
[5] Benjamini Y, Liu W. A distribution-free multiple test procedure that controls the false discovery rate. Tel Aviv. RP-SOR-99-3: Department of Statistics and O.R., Tel Aviv University, 1999.
[6] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependancy. The Annals of Statistics, 2001 (In press).
[7] Crabbe JC, Wahlsten D, Dudek BC. Genetics of mouse behavior: interactions with laboratory environment. Science 1999;284:1670–2.
[8] Drai D, Kafkafi N, Benjamini Y, Elmer G, Golani I. Rats and mice share common ethologically relevant parameters of Exploratory behavior, Behav Brain Res 2001; (Conditionally Accepted to this Volume).
[9] Hochberg H, Tamhane A. Multiple Comparisons Procedures. NewYork: John Wiley & Sons, 1987.
[10] Hsu JC. Multiple Comparisons. London: Chapman and Hall, 1996.
[11] Keselman HJ, Cribbie R, Holland B. The pairwise multiple comparison multiplicity problem: an alternative approach to familywise and comparisonwise type I error control. Psychol Methods 1999;4(1):58–69.
[12] Mallet L, Mazoyer B, Martinot JL. Functional connectivity in depressive, obsessive-compulsive, and schizophrenic disorders: an explorative correlational analysis of regional cerebral metabolism. Psychiatr Res Neuroimaging 1998;82(2):83–93.
[13] Troendle JF. Stepwise normal theory multiple test procedures controlling the false discovery rate. J Stat Plann Inference 2000;84(1-2):139–58.
[14] Weller JI, Song JZ, Heyen DW, Lewin HA, Ron M. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. Genetics 1998;150(4):1699–706.
[15] Westfall PH, Young C. Resampling Based Multiple Comparison Procedures. Wiley-interscience, 1993.
[16] Williams VSL, Jones LV, Tukey JW. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. J Educ Behav Stat 1999;24(1):42–69.
[17] Yekutieli D, Benjamini Y. Resampling based false discovery rate controlling procedure for dependent test statistics. J Stat Plann Inference 1999;82(1-2):171–90.