Conservation Genetics (2019) 20:927–937
https://doi.org/10.1007/s10592-019-01178-0

SHORT COMMUNICATION

Beyond Bonferroni revisited: concerns over inflated false positive research findings in the fields of conservation genetics, biology, and medicine

Tonya White[1,2] · **Jan van der Ende**[1] · **Thomas E. Nichols**[3,4,5]

Received: 5 September 2018 / Accepted: 25 March 2019 / Published online: 11 April 2019

Abstract

In 2006, Narum published a paper in Conservation Genetics emphasizing that Bonferroni correction for multiple testing can be highly conservative with poor statistical power (high Type II error). He pointed out that other approaches for multiple testing correction can control the false discovery rate (FDR) with a better balance of Type I and Type II errors and suggested that the approach of Benjamini and Yekutieli (BY) 2001 provides the most biologically relevant correction for evaluating the significance of population differentiation in conservation genetics. However, there are crucial differences between the original Benjamini and Yekutieli procedure and that described by Narum. After carefully reviewing both papers, we found an error due to the incorrect implementation of the BY procedure in Narum (Conserv Genet 7:783–787, 2006) such that the approach does not adequately control FDR. Since the incorrect BY approach has been increasingly used, not only in conservation genetics, but also in medicine and biology, it is important that the error is made known to the scientific community. In addition, we provide an overview of FDR approaches for multiple testing correction and encourage authors first and foremost to provide effect sizes for their results; and second, to be transparent in their descriptions of multiple testing correction. Finally, the impact of this error on conservation genetics and other fields will be study-dependent, as it is related to the number of true to false positives for each study.

Keywords Multiple testing correction · False discovery rate · Family-wise error · Benjamini Hochberg · Benjamini Yekutieli

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10592-019-01178-0) contains supplementary material, which is available to authorized users.

✉ Tonya White
t.white@erasmusmc.nl

[1] Department of Child and Adolescent Psychiatry, Erasmus University Medical Center, Erasmus MC-Sophia/Kamer KP-2869, Postbus 2060, 3000 CB Rotterdam, The Netherlands

[2] Department of Radiology, Erasmus University Medical Center, Rotterdam, The Netherlands

[3] Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK

[4] Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, UK

[5] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

Introduction

In 2006, Narum published a paper in Conservation Genetics pointing out the conservative nature of the Bonferroni approach to correct for multiple testing when considering a set of statistical inferences and the potential for higher Type II errors (Narum 2006). He suggested that alternative approaches, such as the use of false discovery rate (FDR) to correct for multiple testing can be very effective and can provide a better balance between Type I and Type II errors (Type I error is a false positive, incorrectly rejecting a true null hypothesis; whereas Type II error is a false negative, a failure to reject a false null hypothesis). Further, Narum (2006) argued that tests to correct for multiple testing should be chosen on a case-by-case basis depending on the priority of potential Type I and Type II errors. Finally, he proposed the FDR approach of (Benjamini and Yekutieli 2001) as an alternative approach and potentially more biologically relevant for conservation genetics.

His paper, "Beyond Bonferroni: Less conservative analyses for Conservation Genetics," has been cited over 600 times to date. The article has not only been cited in the field of conservation genetics, but also has been increasingly cited in the fields of biology and medicine. These studies apply the equation described by Narum (2006) attributed to the Benjamini and Yekutieli (2001) procedure for multiple testing correction (BY-FDR). However, a careful review of the published BY method and what Narum describes as the BY method shows crucial differences. Close examination of the two works shows that not all steps were included in calculating the BY-FDR procedure in Narum (2006), and thus this implementation of BY is incorrect and cannot be guaranteed to control the FDR. Thus, we believe that this error has created confusion about the BY procedure and the misimplementation is being propagated along an increasing number of studies.

Within this context, we have three goals of this paper: The first is to provide an overview of the Bonferroni method, the original (Benjamini and Hochberg 1995) FDR (BH-FDR), and the Benjamini and Yekutieli (2001) method (BY-FDR); the second goal is to describe the incorrect implementation of the BY-FDR approach described by Narum, which we will henceforth label as the BY-mis (short for BY-Misimplementation) approach; and the third is to assess the potential impact of this error using 30 of the most recent publications that cite the Narum (2006) paper. However, with the large number of papers that have applied this approach, the specific impact within the fields of conservation genetics, biology, and medicine will need to be evaluated by experts within each of the domains or sub-domains of research in these fields. We will demonstrate that using the BY-mis approach for multiple testing correction results in higher rates of false positives, especially when a large number of multiple tests are performed. However, as pointed out by Narum (2006), false negatives can also be a concern and specific situations may require approaches that limit Type II errors. Typically larger sample sizes are needed to confirm true negatives. In situations where sample sizes are low, as is often the case in conservation genetics (e.g., low number of sampled individuals and/or populations, low number of loci in non-model species) decisions based on false negatives could lead to less productive conservation management strategies (Narum 2006). Thus, we also provide simulations to demonstrate the rates of false negatives using different approaches for multiple testing correction in two specific scenarios.

Theory

We first review the different multiple testing approaches discussed by Narum (2006) using his notation as closely as possible. We start with a collection of k tests, each with a corresponding p value, p_i, $i=1,\ldots,k$. A multiple testing procedure identifies a subset of the k tests as significant while controlling some measure of false positive risk that takes into account the number of tests performed. The Bonferroni method controls the family-wise error (FWE), the chance of one or more false positives, by using a fixed threshold of:

$$\alpha_{\text{Bonf}} = \frac{1}{k}\alpha_{\text{FWE}}$$

where α_{FWE} is the desired FWE level: All tests with $p_i \leq \alpha_{\text{Bonf}}$ can be declared significant while controlling the FWE.

Benjamini and Hochberg (1995) introduced the false discovery rate (FDR) for multiple testing correction. In describing the FDR it is useful to first define the false discovery proportion (FDP): FDP is the ratio of the number of false positive tests to total number of significant tests, defined as 0 if no tests are significant. The FDR is the expected value of FDP; put another way, FDR is the expected proportion of false positives among positives. To find FDR-significant tests, denote the ordered p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$. Then for a desired α_{FDR}, let the index i^* be found as

$$i^* = \max\left\{ i : p_{(i)} \leq \frac{i}{k}\alpha_{\text{FDR}} \right\},$$

and the tests with $p_i \leq p_{(i^*)}$ can be declared significant while controlling FDR at α_{FDR}.

The assumptions of this BH-FDR procedure (BH-FDR) are independence among the test statistics (Benjamini and Hochberg 1995). However, Benjamini and Yekutieli (2001) found that weaker assumptions could be used, allowing a general form of positive dependence among the test statistics. They proposed another method for controlling FDR that makes *no assumptions* about the dependence among the tests, as long as a more stringent criterion was used (Theorem 1.3, BY), with the index i^*_{BY} computed:

$$i^*_{\text{BY}} = \max\left\{ i : p_{(i)} \leq \frac{i}{k}\frac{1}{\sum_{i'=1}^{k}\frac{1}{i'}}\alpha_{\text{FDR}} \right\}.$$

With this approach, the tests with $p_i \leq p_{(i^*_{\text{BY}})}$ are marked significant and FDR is controlled at α_{FDR} under any form of dependency. Note that $\sum_{i'=1}^{k}\frac{1}{i'} \approx \log(k) + \gamma$, where $\gamma \approx 0.57721$ is Euler–Mascheroni constant. This is the method we refer to by BY-FDR.

We can now make a quick comparison of three methods on the basis of the smallest p-value $p_{(1)}$: Bonferroni has the fixed threshold α_{FWE}/k, while BH-FDR will compare $p_{(1)}$ to α_{FDR}/k and BY-FDR will compare $p_{(1)}$ to approximately $\alpha_{\text{FDR}}/(k \log(k))$. Of course, BH-FDR and BY-FDR are adaptive and thus the comparison for each p-value within a test set has successively more lenient thresholds. However, as BH-FDR and BY-FDR use the same inequality except for

the $\approx 1/\log(k)$ term, BY-FDR can only be more stringent than BH-FDR.Now, in Narum (2006), the author incorrectly states that the BY-FDR threshold is fixed and equal to:

$$\frac{1}{\sum_{i=1}^{k}\frac{1}{i}}\alpha_{\mathrm{FDR}}.$$

This is a fundamental error, as a key feature of FDR methods is that they are adaptive. The error arose from neglecting that this expression was just one component of the BY procedure [to be substituted for q in BY Eq. (1) on pp. 1167 (Benjamini and Yekutieli 2001)]. This incorrect application of the BY approach (BY-mis) results in a fixed threshold for a specific k.

Since a fixed threshold specifies the average or per comparison error rate (PCE), we have taken several approaches to assess the impact of this error. Assuming the complete null, i.e. no signal for any test, $k \times \mathrm{PCE}$ is the expected number of false positives. For the threshold at the 0.05 level, for $k = 105$, BY-mis has $k \times \mathrm{PCE} \approx 1$, while for $k = 1590$, $k \times \mathrm{PCE} \approx 10$. This demonstrates that the BY-mis approach can be assured to produce an increasing number of false positives for an increasing k. In contrast, for Bonferroni $k \times \mathrm{PCE}$ is exactly α_{FWE}, i.e. always less than 1, and every valid FWE or FDR level α procedure is guaranteed to produce no false positives with probability 1−α (again, in this complete null setting). While the BY-mis approach does asymptote to zero as k approaches infinity, it approaches zero extremely slowly. For example, with 10 million tests performed, the BY-mis p-value threshold is 0.003, in contrast to the Bonferroni threshold of 0.000000005.

To evaluate the rate of significant p-values found with the Bonferroni, BH, BY, BY-mis, and uncorrected approaches we conducted a simulation using the Python programming language version 2.7.13 (Zope Corporation and a cast of thousands;www.python.org); the code used for all simulations is available in the supplement.. We performed simulations using k values ranging from 1 to 100 tests. For each k, we created 50,000 random realizations where null p-values were computed from test statistics generated as a standard normal distribution. Thus, for $k = 1$ we had a total of 50,000 independent p-values and in this case the four approaches were identical. For $k > 1$ we generated k independent p-values and applied each of the four methods. A nominal $\alpha_{\mathrm{FWE}} = \alpha_{\mathrm{FDR}} = 0.05$ was used for all methods. In this null setting, any "discovery" is a false discovery and so the measured FDR and FWE are the same. We computed the proportion of realizations where any p-values were found significant, representing a FWE error and a FDP of 1. Figure 1a shows the FDR and FWE as a function of the number of tests, showing that Bonferroni and BH-FDR both control false positives as expected (as an aside, while Bonferroni is often regarded as conservative, in this setting of small k and independent tests, it is essentially exact). The FDR/FWE of BY-FDR becomes increasing conservative while the BY-mis has inflated false positives with a near linear increase with increasing k.

In addition, we performed simulations using python to measure both false negative rates for the Bonferroni, BH, BY, and the BY-mis approaches for multiple testing correction. These simulations were creating 50,000 realizations of sets of k tests, 1 to 100, but in this simulation we included a mix of null and non-null tests. We performed two classes of simulations, one with 1 non-null test and one with 25 non-null tests. For example, with $k = 50$ and the situation of 1 non-null test, there were 49 random p-values computed from a standard Normal distribution test statistic, and 1 p-value that was generated with from a non-null Normal with mean set to give a test with 80% power at the uncorrected level $\alpha = 0.05$. The same situation with $k = 50$ for the case with 25 non-null tests, where 25 p-values were generated from null test statistics and 25 non-null p-values were generated to have 80% power to reject the null. This can be seen in Fig. 1b, c where the probability of a false negative for uncorrected comparisons remains at 0.2. These simulations show that the BY-FDR, has the highest probability of a Type II error with one simulated non-null result, whereas the BH-FDR and Bonferroni are very similar.

To illustrate these simulations with an example, say that a study was conducted in which 50 tests were performed ($k = 50$) with half of the tests actually being significant. Thus, there are 25 tests in which there is a possibility of false positive, and 25 tests in which there is a possibility of a false negative. Since Fig. 1c models the case of 25 out of $k = 25$ to 100 significant tests, the probability of a false negative for $k = 50$ is approximately 0.4 for the BH-FDR, 0.43 for the BY-mis, 0.68 for the BY-FDR, and 0.72 for the Bonferroni approach. The probability of a false positive for 25 non-significant tests can be determined from Fig. 1a. With $k = 25$, the FDR and FWE rate would be at approximately 5% and below for the Bonferroni, BH-FDR, and BY-FDR, but the false discovery rate would be approximately 30% for the BY-mis (Fig. 1a). Figure 1b, c shows that for all methods used to correct for multiple testing, the risk of Type II error increases with the number of tests k. However, there is a dramatic difference between the performance of BY-FDR and the BY-mis. Note the advantage of the BH-FDR approach in minimizing both false positive and false negative errors, while still controlling FDR.

We also consider the specific set of 15 p-values used in Narum (2006) to tabulate the p-value thresholds for the Bonferroni, BH, BY, and the BY-mis approaches. Table 1 shows the thresholds used for each of the 15-exemplar p-values, with significant tests marked in bold. It can be seen that the BY-FDR and the BY-mis are not the same. Narum (2006) reported four significant tests as compared to the correct BY-FDR's having two significant tests.

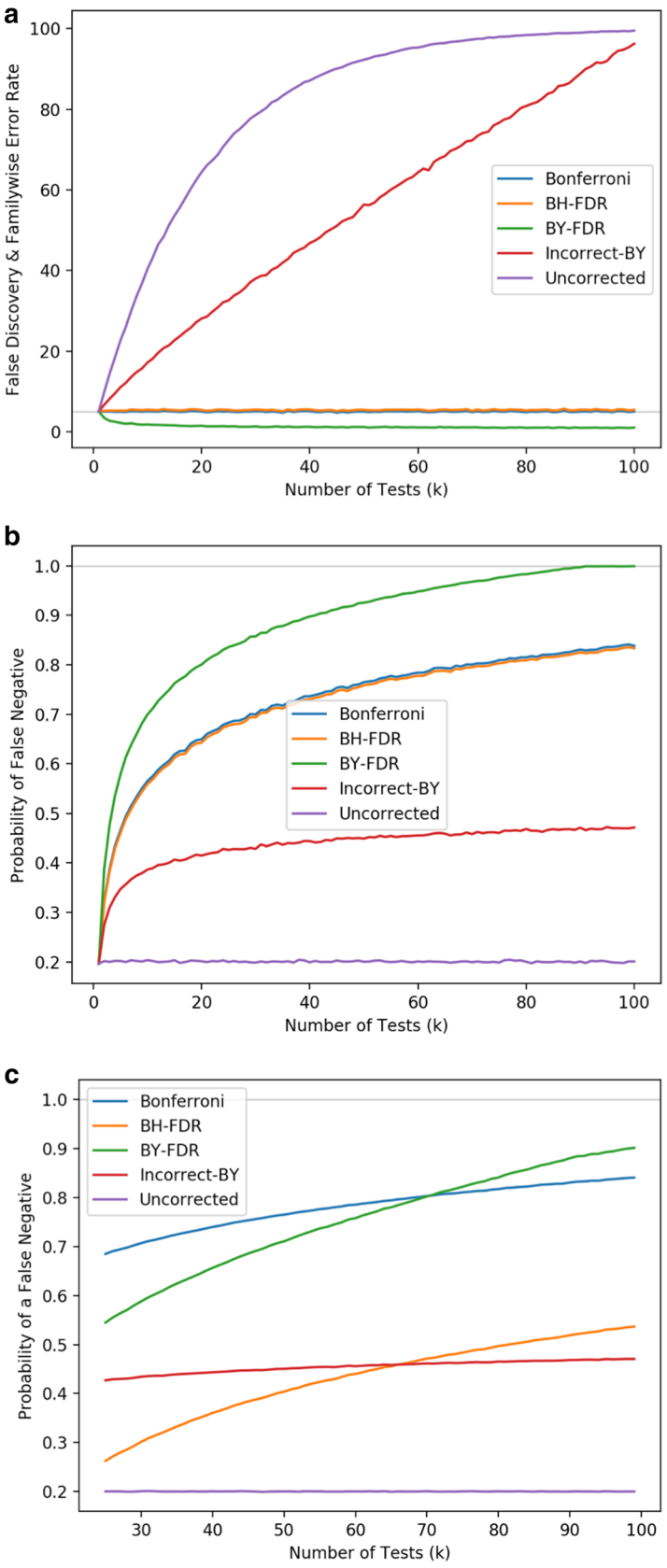

Fig. 1 Probability of Type I and Type II errors compared to the number of independent tests performed. **a** False positive rates under the complete null setting, showing false discovery and family-wise error rate (here, identical) plotted against the number of tests performed using five different approaches: Bonferroni, Benjamini–Hochberg (BH-FDR), Benjamini and Yekutieli (BY-FDR), the BY-misimplementation (BY-mis), and no correction. It is demonstrated in this simulation that the FDR and FWE rise dramatically with k (the number of tests) for BY-mis. **b** Type II error rates for a one non-null test out of a total of k tests (k = 1–100). **c** Average Type II error rate over 25 non-null tests out of k tests (k = 25–100). Type II error rates rise with k for all multiple testing methods, but BY-mis has dramatically different rates than BY-FDR. A total of 50,000 iterations were done for each simulation and the python code is provided in the supplement

Table 1 A set of p-values from 15 significance testing taken from the Narum 2006 paper (column labeled 'p-value examples') and comparison with four approaches to multiple testing (critical p-values for significance)

p-value examples	Bonferroni	Benjamini and Hochberg	Benjamini and Yekutieli	BY-misimplementation
0.0001	**0.0033**	**0.0033**	**0.0010**	**0.0151**
0.0010	**0.0033**	**0.0067**	**0.0020**	**0.0151**
0.0062	0.0033	**0.0100**	0.0030	**0.0151**
0.0101	0.0033	**0.0133**	0.0040	**0.0151**
0.0214	0.0033	**0.0167**	0.0050	0.0151
0.0227	0.0033	**0.0200**	0.0060	0.0151
0.0273	0.0033	**0.0233**	0.0070	0.0151
0.0292	0.0033	**0.0267**	0.0080	0.0151
0.0311	0.0033	**0.0300**	0.0090	0.0151
0.0323	0.0033	**0.0333**	0.0100	0.0151
0.0441	0.0033	0.0367	0.0111	0.0151
0.0490	0.0033	0.0400	0.0121	0.0151
0.0573	0.0033	0.0433	0.0131	0.0151
0.1262	0.0033	0.0467	0.0141	0.0151
0.5794	0.0033	0.0500	0.0151	0.0151

Numbers in bold reflect the 'p-value examples' that are significant based on each of the four critical p-value columns

The example in Table 1 also demonstrates one of the challenges in finding a balance between Type I and Type II errors and the choice for multiple testing correction. The probability that 12 of 15 independent tests would show an uncorrected p-value less than 0.05 is very low. Thus, Bonferroni, having only two significant tests, is likely overly conservative and would result in a higher type II error rate. The BH-FDR approach, however, shows that 10 of the 15 tests are identified as significant, which in this situation may be more plausible, although it would be helpful to know the covariance structure between the different variables, as statistical dependence between variables is not uncommon. Figure 1c demonstrates type II error rates for the simulation of 25 true positives (80% chance of being less than $p < 0.05$) and the notable differences between the Bonferroni and BH-FDR for $k = 1–100$ independent tests.

Finally, we used Scopus to identify the 30 most recent publications (search date: February, 9, 2019) that cite Narum (2006) to sample the impact of this error on the literature (Table 2). Of these 30 articles, nine articles (30%) were specifically related to conservation genetics; ten articles were in the fields of biology, mostly involving genetic analyses (33.3%); nine articles (30%) were in the field of medicine, most commonly in psychiatry; and the two additional articles were in the fields of statistics and anthropology. In 20 of these articles (67%) we could confidently determine that BY-mis was used (2006), while it was unclear in six articles (20%), and one article cited Narum (2006), but did not use the BY-mis approach. None of the papers described using a standard statistical software package to calculate the BY-FDR. Eight of the twenty articles that applied the BY-mis approach also cited the Benjamini and Yekutieli (2001) article. Of the 28 relevant articles [excluding Hauser et al. (2018) and Stepien et al. (2018) as these papers cited but did not apply the BY-mis approach], only eight articles (29%) provide enough information to calculate the alternate multiple testing corrections for the data provided for the specific study. Four of these eight articles show an reduction in the number of significant tests when BY-mis is replaced with BY-FDR, whereas the other four have tests that either are negative (one article) or are so strongly significant that all the tests also pass Bonferroni correction (three articles). Also noteworthy, eight of the twenty articles that applied the BY-mis approach (40%) applied independent levels of multiple testing, rather than applying multiple testing to all tests in the article.

Discussion

In 1995, Benjamini and Hochberg proposed the FDR metric and a method to control FDR. Benjamini and Yekutieli in 2001 proposed a method to control FDR with weaker assumptions, but more stringent correction than the BH approach. Narum's (2006) paper provided an overview and examples of the BY-FDR procedure, however, did not include all steps of the BY algorithm (shown above). A careful reading of Benjamini and Yekutieli (2001) reveals that the equation for multiple testing from Narum (2006) (from Theorem 1.3 on pp. 1169 of BY) should be entered as the α in the B-H equation (Eq. (1) on pp. 1167 in BY), producing an adaptive threshold. Further, based on a series of p-values taken from the Narum (2006) paper (Table 1), different results are obtained comparing the Narum (2006) description of the BY approach and the BY-FDR described by Benjamini and Yekutieli (2001).

Direct calculation shows that BY-mis has expected number of false positives that increases nearly linearly with number of tests *k,* and that this increasing false positive rate differs dramatically from the BY-FDR approach (Fig. 1a). We believe that a large percentage of the over 600 publications are liable to have this inflated rate of false positives in their results, notably since results arising from Type I errors are much easier to publish than those from Type II errors. We found that at least 40% of a sample of the 30 most recent papers that cite Narum (2006) article also cite Benjamini and Yekutieli (2001) and that they have applied the BY approach, but actually apply the BY-mis-FDR approach (Table 2).

Table 2 List of the 30 most recent articles identified via Scopus (9 February 2019) who cited the Narum (2006) article

Reference	Fields of study	Also cited original B-Y paper	Applied critical p-value as described by Narum (2006)	Enough information provided to calculate equations for multiple testing	Total number of tests	Number of significant tests using incorrect B-Y	Number of significant tests using B-H	Number of significant tests using B-Y	Number of significant tests using Bonferroni	Error in multiple testing correction
Xue et al. (2019)	Conservation Genetics	No	Yes	No	144[a]	135	?	?	?	Yes
Paans et al. (2019)	Medicine/Nutrition	No	Yes	No	24	11	?	?	?	Yes
Riesgo et al. (2019, Table 3)	Evolutionary Biology	Yes	Yes	Yes	9[a]	9	9	9	9	Yes[c]
Buchanan et al. (2019, Table 3)	Anthropology	Yes	Yes	Yes	4[a]	1	0	0	0	Yes
Hauser et al. (2018)	Statistics	Yes	n/a	n/a					?	n/a
Sandoval Laurrabaquio-A et al. (2019)	Conservation Genetics	Yes	Yes	No	6[a,b]	?	?	?	?	Yes
Sucec et al. (2019, Figure 1)	Medicine/Psychiatry	Yes	Yes	No	3[a,b]	?	?	?	?	
Deane et al. (2018, Table 2)	Medicine/Psychiatry	No	Yes	No	68	6	?	?	?	Yes
Huang et al. (2018, Table 2)	Conservation Genetics	No	Yes	No	21[a,b]	13	?	?	?	Yes
Austin et al. (2018)	Biology	No	Yes	No	?	?	?	?	?	?
Van Wyk et al. (2018)	Biology	No	Yes	No	?	3	?	?	?	?
Pérez-Portela et al. (2018, Table 4)	Biology	No	Yes	No	54[a,b]	8	?	?	?	Yes
DiBattista et al. (2018, Supplemental Table A.1)	Biology	No	Yes	Yes	45	21	21	21	21	Yes[c]
Wieman and Berendzen (2018, Table 2)	Conservation Genetics	Yes	Yes	No	25[b]	18	?	?	?	Yes

Table 2 (continued)

Reference	Fields of study	Also cited original B-Y paper	Applied critical p-value as described by Narum (2006)	Enough information provided to calculate equations for multiple testing	Total number of tests	Number of significant tests using incorrect B-Y	Number of significant tests using B-H	Number of significant tests using B-Y	Number of significant tests using Bonferroni	Error in multiple testing correction
Pérez-Portela et al. (2019, Supplemental Table S2)	Biology/Marine Sciences	No	Yes	No	78[b]	52	?	?	?	Yes
Pregler et al. (2018, p. 1943)	Conservation Genetics	Yes	Yes	No	37[b]	1	?	?	?	Yes
Hoffmann et al. (2018, Supplemental Table S2)	Medicine/Psychiatry	No	Yes[f]	Yes	5	1	0	0	0	Yes
Bartholomeusz et al. (2018, Table 5)	Medicine/Psychiatry	No[e]	Yes	Yes	12	0	0	0	0	No
Gibson-Smith et al. (2018, p. 4)	Medicine/Psychiatry	No	?	No	30	15	?	?	?	Yes
Davis et al. (2018, p. 42)	Biology/marine sciences	No	?[g,i]	No	?	?	?	?	?	?
Barendse et al. (2018, Table 3)	Medicine/Psychiatry	No	?	No	16[b]	0	?	?	?	?
Sucec et al. (2018, p. 44)	Medicine/psychophysiology	Yes	Yes	Yes	4[a,b]	4	4	4	4	No[h]
Xue et al. (2018, Table 2)	Conservation Genetics	No	Yes	No	?	?	?	?	?	Yes
Hasselman et al. (2018, Supplemental Table S3)	Conservation Genetics	Yes	?	Yes/No	66	61	?	?	?	No
Casey et al. (2018, no page numbers)	Conservation Genetics	Yes	?	No	?	?	?	?	?	Yes
Petereit et al. (2018, p. 1128)	Conservation Genetics	Yes	?	No	276	151	153	?	?	Yes

Table 2 (continued)

Reference	Fields of study	Also cited original B-Y paper	Applied critical p-value as described by Narum (2006)	Enough information provided to calculate equations for multiple testing	Total number of tests	Number of significant tests using incorrect B-Y	Number of significant tests using B-H	Number of significant tests using B-Y	Number of significant tests using Bonferroni	Error in multiple testing correction
McDowell et al. (2018, Table 2)	Biology/Marine Sciences	Yes	Yes	Yes	14	1	0	0	0	Yes
Mutton et al. (2018, Table 3)	Ecology and Evolution	No	Probably yes	No	21	21	?	?	?	Yes
Stepien et al. (2018, p. 787)	Biology/Marine Sciences	No	[d]	No						
Mahdavi et al. (2018, Table 3)	Medicine/Proteomics	No	Yes	Yes	54	4[j]	0	0	0	Yes

Information is provided on the articles and, if enough information is provided, the calculations for multiple testing correction for the Bonferroni, B-H, B-Y, and the Incorrect B-Y

[a]There were groups of analyses performed in which the multiple testing corrections were applied independently to the different groups. We present either the first group encountered or alternatively, the first group that provides adequate information to assess the multiple testing correction

[b]The total number of tests field was calculated from the critical p-value provided in the manuscript by using the equation from Narum (2006)

[c]The misimplementation of Narum (2006) was applied, but all findings were highly significant and thus the error did not result in any differences between the types of FDR

[d]Cited Narum but did not use the correction for multiple testing as described by Narum (2006)

[e]Cited a different paper by Benjamini–Yekatuli that did not present the B-Y equation

[f]Described that they used the Narum (2006) approach, but listed the incorrect equation than that described in Narum (2006)

[g]The adjustment for multiple testing was not mentioned in the results section

[h]An error was likely present but in a different comparison

[i]Applied adjustment for multiple testing but did not specifically report the type of testing

[j]Difference between the number of significant findings reported in the text versus those presented in the listed table

We do agree with Narum that the Bonferroni approach can be highly conservative in some situations of multiple testing correction, especially with dependent data. However, there has also been a growing concern that many studies fail to replicate (Ioannidis 2005; Open Science Collaboration 2015; Nichols et al. 2017; Gelman 2018). In the past, analyses were performed without adequately controlling for the numbers of tests performed (Carp 2012) which resulted in numerous Type I errors but also likely fewer Type II errors. We also agree with Narum that the individual studies should determine the balance between Type I and Type II errors, as there are some situations in many fields where researchers want to limit Type II errors. Examples include situations in conservation genetics where a failure to show a positive effect could direct conservation management strategies that are counter to the survival of a species (Narum 2006). Species in which there is concern over extinction often have smaller populations and lower rates of reproduction (Lynch and Lande 1998) and decisions based on false negatives in some populations could lead to less productive conservation management strategies. Examples in medicine with concerns over false negatives include the presurgical use of functional magnetic resonance imaging to identify eloquent cortex (Durnez et al. 2013). In such cases, a false negative could result in the removal of eloquent cortical regions and thus stringent correction for multiple testing would not be indicated. Thus, in conservation genetics, biology, medicine, and other fields, individual studies may shift the choice of limiting either Type I or Type II errors and providing the rationale for the choice of (or lack of) multiple testing correction should always be provided.

Our attempt to extract vital information to assess the multiple testing correction within each of the 30 most recent articles that cite the Narum (2006) paper highlights the need in the literature for greater transparency regarding the use of multiple testing correction. Over two-thirds of these papers did not provide enough information to replicate the authors approach for multiple testing correction nor to compare the different methods. Further, a minority of these papers presented effect sizes or confidence intervals for their findings, and omission of these data been shown to be a problem in many fields of science (Chavalarias et al. 2016). None of the authors described using statistical software packages, (i.e. R or SAS) to calculate the BY-FDR, which, if performed correctly, would have resulted in an accurate calculation of multiple testing correction. It is likely that the BY-mis approach, which provides a single critical p-value and is trivial to calculate, was easier than the use of statistical software. There is currently discussions regarding moving away from the use of the $p \leq 0.05$ approach (American Statistical Association 2016), we would recommend that if p-values are presented, they should always be the full, unadjusted p-value and should be accompanied by effect sizes or confidence intervals. Effect sizes or confidence intervals provide greater details regarding hypothesis testing compared to p-values (Smith 2018) and will enhance replication, as studies evaluating small effects in the wake of considerable noise are likely false positives (Gelman 2018), considering a system rewarded by positive findings.

In summary, so long as p-values remain one of the top methods of choice to report statistical results, we agree with the Narum (2006) that researchers should carefully consider the different tests for multiple testing correction and should make a priori decisions based on Type I and Type II errors within their specific study. Further, we provide an overview of FWE and FDR correction approaches and several simulations to show both type I and type II errors. We point out an error in Narum's (2006) paper describing the BY approach and show that the BY-mis does not adequately control for FDR when used for multiple testing correction. Finally, we recommend that authors be transparent in reporting the number of tests, the number of clusters of tests, and method used when performing multiple testing correction. Authors should also present effect sizes or confidence intervals is also key.

Funding Funding was provided by ZonMw (Grant No: 91211021).

References

American Statistical Association (2016) American Statistical Association releases statement on statistical significance and p -values: provides principles to improve the conduct and interpretation of quantitative science. ASA News. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN. Accessed 7 Mar 2019

Austin JD, Greene DU, Honeycutt RL, McCleery RA (2018) Genetic evidence indicates ecological divergence rather than geographic barriers structure florida fox squirrels. J Mammal. https://doi.org/10.1093/jmammal/gyy128

Barendse MEA, Simmons JG, Byrne ML et al (2018) Associations between adrenarcheal hormones, amygdala functional connectivity and anxiety symptoms in children. Psychoneuroendocrinology 97:156–163. https://doi.org/10.1016/j.psyneuen.2018.07.020

Bartholomeusz CF, Ganella EP, Whittle S et al (2018) An fMRI study of theory of mind in individuals with first episode psychosis. Psychiatry Res Neuroimaging 281:1–11. https://doi.org/10.1016/j.pscychresns.2018.08.011

Benjamini Y, 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, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188. https://doi.org/10.2307/2674075

Buchanan B, Hamilton MJ, Hartley JC, Kuhn SL (2019) Investigating the scale of prehistoric social networks using culture, language, and point types in western North America. Archaeol Anthropol Sci 11:199–207. https://doi.org/10.1007/s12520-017-0537-y

Carp J (2012) The secret lives of experiments: methods reporting in the fMRI literature. Neuroimage 63:289–300. https://doi.org/10.1016/j.neuroimage.2012.07.004

Casey CS, Orozco-terWengel P, Yaya K et al (2018) Comparing genetic diversity and demographic history in co-distributed wild South American camelids. Heredity (Edinb) 121:387–400. https://doi.org/10.1038/s41437-018-0120-z

Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA (2016) Evolution of reporting *P* values in the biomedical literature, 1990-2015. JAMA 315:1141. https://doi.org/10.1001/jama.2016.1952

Davis AR, Becerro M, Turon X (2018) Living on the edge: early life history phases as determinants of distribution in *Pyura praeputialis* (Heller, 1878), a rocky shore ecosystem engineer. Mar Environ Res 142:40–47. https://doi.org/10.1016/j.marenvres.2018.09.019

Deane C, Vijayakumar N, Allen NB et al (2018) Parentingxbrain development interactions as predictors of adolescent depressive symptoms and well-being: differential susceptibility or diathesis-stress? Dev Psychopathol. https://doi.org/10.1017/s0954579418001475

DiBattista JD, Wakefield CB, Moore GI et al (2018) Genomic and life-history discontinuity reveals a precinctive lineage for a deep-water grouper with gene flow from tropical to temperate waters on the west coast of Australia. Ecol Genet Genomics 9:23–33. https://doi.org/10.1016/j.egg.2018.09.001

Durnez J, Moerkerke B, Bartsch A, Nichols TE (2013) Alternative-based thresholding with application to presurgical fMRI. Cogn Affect Behav Neurosci 13:703–713. https://doi.org/10.3758/s13415-013-0185-3

Gelman A (2018) The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. Personal Soc Psychol Bull 44:16–23. https://doi.org/10.1177/0146167217729162

Gibson-Smith D, Bot M, Brouwer IA et al (2018) Diet quality in persons with and without depressive and anxiety disorders. J Psychiatr Res 106:1–7. https://doi.org/10.1016/j.jpsychires.2018.09.006

Hasselman DJ, Bentzen P, Narum SR, Quinn TP (2018) Formation of population genetic structure following the introduction and establishment of non-native American shad (*Alosa sapidissima*) along the Pacific Coast of North America. Biol Invasions 20:3123–3143. https://doi.org/10.1007/s10530-018-1763-7

Hauser S, Wakeland K, Leberg P (2018) Inconsistent use of multiple comparison corrections in studies of population genetic structure: Are some type I errors more tolerable than others? Mol Ecol Resour. 19:144–148

Hoffmann C, Van Rheenen TE, Mancuso SG et al (2018) Exploring the moderating effects of dopaminergic polymorphisms and childhood adversity on brain morphology in schizophrenia-spectrum disorders. Psychiatr Res Neuroimaging 281:61–68. https://doi.org/10.1016/j.pscychresns.2018.09.002

Huang W, Li M, Yu K et al (2018) Genetic diversity and large-scale connectivity of the scleractinian coral *Porites lutea* in the South China Sea. Coral Reefs 37:1259–1271. https://doi.org/10.1007/s00338-018-1724-8

Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2:0696–0701. https://doi.org/10.1371/journal.pmed.0020124

Lynch M, Lande R (1998) The critical effective size for a genetically secure population. Anim Conserv 1:70–72

Mahdavi S, Jenkins DJA, Borchers CH, El-Sohemy A (2018) Genetic variation in 9p21 and the plasma proteome. J Proteome Res 17:2649–2656. https://doi.org/10.1021/acs.jproteome.8b00117

McDowell JR, Mamoozadeh NR, Brightman HL, Graves JE (2018) Use of rapidly evolving molecular markers to distinguish species and clarify range uncertainties in the spearfishes (Istiophoridae, *Tetrapturus*). Bull Mar Sci 94:1355–1378. https://doi.org/10.5343/bms.2017.1130

Mutton TY, Fuller SJ, Tucker D, Baker AM (2018) Discovered and disappearing? Conservation genetics of a recently named Australian carnivorous marsupial. Ecol Evol 8:9413–9425. https://doi.org/10.1002/ece3.4376

Narum SR (2006) Beyond Bonferroni: less conservative analyses for conservation genetics. Conserv Genet 7:783–787

Nichols T, Das S, Evans AC et al (2017) Best practices in data analysis and sharing in neuroimaging using MRI best practices in data analysis and sharing in neuroimaging using MRI. Nat Neurosci 20:299–303. https://doi.org/10.1038/nn.4500

Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349:aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Paans NPG, Gibson-Smith D, Bot M et al (2019) Depression and eating styles are independently associated with dietary intake. Appetite 134:103–110. https://doi.org/10.1016/j.appet.2018.12.030

Pérez-Portela R, Bumford A, Coffman B et al (2018) Genetic homogeneity of the invasive lionfish across the Northwestern Atlantic and the Gulf of Mexico based on single nucleotide polymorphisms. Sci Rep 8:5062. https://doi.org/10.1038/s41598-018-23339-w

Pérez-Portela R, Wangensteen OS, Garcia-Cisneros A et al (2019) Spatio-temporal patterns of genetic variation in *Arbacia lixula*, a thermophilous sea urchin in expansion in the mediterranean. Heredity (Edinb) 122:244–259. https://doi.org/10.1038/s41437-018-0098-6

Petereit C, Bekkevold D, Nickel S et al (2018) Population genetic structure after 125 years of stocking in sea trout (*Salmo trutta* L.). Conserv Genet 19:1123–1136. https://doi.org/10.1007/s10592-018-1083-6

Pregler KC, Kanno Y, Rankin D et al (2018) Characterizing genetic integrity of rear-edge trout populations in the southern Appalachians. Conserv Genet 19:1487–1503. https://doi.org/10.1007/s10592-018-1116-1

Riesgo A, Taboada S, Pérez-Portela R et al (2019) Genetic diversity, connectivity and gene flow along the distribution of the emblematic Atlanto-Mediterranean sponge *Petrosia ficiformis* (Haplosclerida, Demospongiae). BMC Evol Biol 19:24. https://doi.org/10.1186/s12862-018-1343-6

Sandoval Laurrabaquio-A N, Islas-Villanueva V, Adams DH et al (2019) Genetic evidence for regional philopatry of the Bull shark (*Carcharhinus leucas*), to nursery areas in estuaries of the Gulf of Mexico and western North Atlantic ocean. Fish Res 209:67–74. https://doi.org/10.1016/j.fishres.2018.09.013

Smith RJ (2018) The continuing misuse of null hypothesis significance testing in biological anthropology. Am J Phys Anthropol 166:236–245. https://doi.org/10.1002/ajpa.23399

Stepien CA, Snyder MR, Knight CT (2018) Genetic divergence of nearby walleye spawning groups in central lake erie: implications for management. North Am J Fish Manag 38:783–793. https://doi.org/10.1002/nafm.10176

Sucec J, Herzog M, Van Diest I et al (2018) The impairing effect of dyspnea on response inhibition. Int J Psychophysiol 133:41–49. https://doi.org/10.1016/j.ijpsycho.2018.08.012

Sucec J, Herzog M, Van Diest I et al (2019) The impact of dyspnea and threat of dyspnea on error processing. Psychophysiology 56:e13278. https://doi.org/10.1111/psyp.13278

Van Wyk AM, Kotzé A, Grobler JP et al (2018) Isolation and characterization of species-specific microsatellite markers for blue and

black wildebeest (*Connochaetes taurinus* and *C. gnou*). J Genet 97:101–109. https://doi.org/10.1007/s12041-018-1000-2

Wieman AC, Berendzen PB (2018) Spatial genetic variation and habitat association of *Rhinichthys cataractae*, the longnose dace, in the driftless area of the upper mississippi River basin. Conserv Genet 19:1367–1378. https://doi.org/10.1007/s10592-018-1106-3

Xue D-X, Graves J, Carranza A et al (2018) Successful worldwide invasion of the veined rapa whelk, *Rapana venosa*, despite a dramatic genetic bottleneck. Biol Invasions 20:3297–3314. https://doi.org/10.1007/s10530-018-1774-4

Xue D-X, Yang Q-L, Li Y-L et al (2019) Comprehensive assessment of population genetic structure of the overexploited Japanese grenadier anchovy (*Coilia nasus*): implications for fisheries management and conservation. Fish Res 213:113–120. https://doi.org/10.1016/j.fishres.2019.01.012