Chapter 17

Bioinformatics of Correction of Multiple Testing: An Introduction for Life Scientists

(in Bioinformatics and Human Genomics Research)

Antonio Carvajal-Rodríguez

Departamento de Bioquímica, Genética e Inmunología y Centro de Investigación Mariña (CIM-UVIGO), Universidade de Vigo, Vigo, Spain

1 INTRODUCTION

Over the past years, biologists are generating data on a massive scale due to technologies such as microarray and high-throughput sequencing. In this biological big data era in which large numbers of hypotheses are simultaneously tested, even on the scale of hundreds of thousands or millions, the multiple testing correction has become of great importance.

Multiple hypothesis testing becomes standard in many scientific areas such as pharmacogenetics, genomics and proteomics. For example, understanding the genetic basis for a certain disease implies testing the expression of thousands of genes between different groups of patients. Thus, the interest for multiple testing procedures (MTPs) and the trade-off between false positive error and statistical power becomes key in current research.

There is a large number of approaches for multiple testing correction. In this review we just outline the basic concepts and strategies. More detailed and complete descriptions are reviewed elsewhere (Goeman and Solari, 2014; Kang, 2020a; Korthauer et al., 2019; MacDonald et al., 2019; Rudra et al., 2019; Tamhane and Gou, 2018). The present chapter is organized in six sections including this Introduction. Section 2 gives an informal overview about the principal multiple testing strategies and their utility under different research scenarios. Section 3 provides

*Corresponding author: *acraaj@uvigo.es*

commented definitions of important concepts. Section 4, provides a more formal and extended technical description, with formulae and algorithms for the different methods described in Section 2. Section 5 briefly, sketches some recent approaches for multiple testing correction. Finally, Section 6 concludes the previous sections.

2 MULTIPLE TESTING CORRECTION OVERVIEW

The scientific method works with proposed explanations for a phenomenon under study. These explanations are called scientific hypotheses provided they can be tested in some way. A test is a rule for deciding whether to accept or reject a hypothesis. The hypothesis under test is called the main or null hypothesis.

Consider an experiment that compares the expression levels of a certain gene in two groups, cases and controls, with sample size $n = 20$ each. According to the null hypothesis, the gene expression is equal between cases and controls. Meanwhile the rejection of the null hypothesis implies that there are differences in the gene expression between cases and controls. Let us consider a test for the hypothesis of equal gene expression, if we perform the test at level α, the risk of rejecting the null hypothesis of equal gene expression when it is true, has a probability of α which is called the type I error probability (Larson, 1982).

Hypotheses testing (Neyman-Pearson sense) consists in applying a rule to a function of the observed data for deciding whether to accept or reject an hypothesis. In applying that rule, the researcher decides in advance the maximum type I error rate (α) that she/he considers acceptable.

In summary, when we perform a single test, two types of errors can occur. The type I error, or false positive, is committed when rejecting a true null hypothesis; the type II error, or false negative, is committed when accepting a false null hypothesis (Table 1). Thus, type I error is controlled by the user of the test by means of the value α decided before performing the test. For example, defining $\alpha = 0.05$ means that type I error of the test is being controlled to have probability equal or less than 0.05.

Table 1 Possible outcomes when testing a null hypothesis H_0

	H_0 True	***H_0 False***
Reject H_0	Type I error	No error
Accept H_0	No error	Type II error

Nowadays, high-throughput experiments involve not one single but thousands of statistical tests. The problem of multiple testing correction consists in that if we compare the expression of thousands of genes and, as in single testing, we fix the type I error to 0.05 for each test, we can be sure that we will have a bunch of false positives (aka false discoveries). Consequently, we need to control type I error for the whole set of tests in some way similarly as the acceptance level α controls type I error for the single test. Of course, in addition to microarray expression studies there are many other scenarios where the same or a similar problem arises, as the genome wide association studies, and in general the comparison issues that raises from the analysis of omics data related to research on therapies, vaccines, diagnostics, etc. Under all these settings we require multiple testing correction to deal with type I error because this kind of error has the risk of disseminating misleading scientific results.

When performing multiple tests there are more than one definition for the type I error rate in the family of tests. We may consider two different definitions. First option considers the distribution of the number (not a proportion) of individual type I errors; an example of this is the family-wise error rate (FWER). The second option considers the distribution of the false discovery proportion (FDP); an example of this is the average of the FDP also called the false discovery rate (FDR).

The family-wise error rate, FWER, is the probability of at least one type I error in the family of tests, that is (see Table 2) FWER = $\Pr(V \geq 1)$, where V is the number of false positives. The FDP is the proportion of type I errors among the rejected hypotheses, i.e., FDP = V/R, with the convention of taking FDP = 0 when $R = 0$ (Dudoit and Laan, 2008). The FDR is the expectation of the FDP but the convention of FDP = 0, when $R = 0$ requires FDR defined as FDR = $\mathrm{E}[V/R \mid R > 0] \times \Pr(R > 0)$.

Table 2 Numbers of Type I and II errors when performing M tests of hypotheses

	H_0 *True*	H_0 *False*	*Total*
Reject H_0	V	U	$R = V + U$
Accept H_0	$M_0 - V$	$M_1 - U$	$M - R$
Total	M_0	M_1	$M = M_0 + M_1$

In Table 2 we may appreciate that if from a total of M hypotheses there are M_0 for which the null is true and M_1 for which is false, then after R rejections we may commit V wrong rejections (type I errors) plus U correct rejections. Similarly, for $M - R$ null acceptances, there are $M_1 - U$ acceptances when the null is false (type II errors) and $M_0 - V$ correct null acceptances.

In this work we focus on three main strategies for managing the type I error under the multiple test setting, namely, those controlling the family-wise error rate (FWER), those controlling the false discovery rate (FDR), and those that estimate the false discovery proportion (FDP). In the following sections an overview is given of different methods based on each strategy, their underlying assumptions jointly with their strengths and weaknesses.

2.1 Methods of Controlling the Family-Wise Error Rate (FWER)

The methods explained below work on a set of M p-values obtained from a family of M tests. A common assumption, except for the permutation based methods, is that the p-values are continuous and uniformly distributed in the (0, 1) interval or they are stochastically larger than uniform in the discrete case (Lehmann and Romano, 2005).

2.1.1 FWER and genomics research

Let us continue with our previous example. Consider that we are comparing the expression levels of 1,000 independent genes for two samples, one with 20 cases and other with 20 controls. For each gene we perform the test under a given critical value (see Section 3.1) that has its proper type I error level, say $\alpha' = 0.05$. In this case, the probability of no type I error for each test is $1 - 0.05 = 0.95$, then for the whole family of 1,000 tests, the probability of committing no error is $(1 - \alpha')^{1,000} = 0.95^{1,000} = 5.3 \times 10^{-23}$ and the probability of at least 1 type I error in the family of tests is FWER = $1 - (1 - \alpha')^{1,000} = 1 - 5.3 \times 10^{-23} \approx 1$.

Therefore, if we perform M tests and want to maintain the type I error below a given rate we need to apply a multiple testing method to guarantee that. Among the many methods available for controlling the type I error rate some of them control directly the FWER (Table 3).

Table 3 FWER control methods with different dependence assumptions and usability in terms of the exploratory or confirmatory nature of the experiments

Method	*Control*	*Dependence assumptions*	*Usability level*	*Software*
Bonferroni	Strong	None	Confirmatory	p.adjust, Myriads v1.2
Holm	Strong	None	Confirmatory	multcomp, Myriads
Hommel	Strong	Positive dependence	Confirmatory	hommel, Myriads v1.2
maxT	Strong	None	Confirmatory	multtest, Myriads v1.2
SGoF	Weak	Independence	Exploratory	sgof, Myriads

2.1.2 Bonferroni

The method of Bonferroni is a single-step procedure that controls FWER at level α by rejecting the set of hypotheses that have p-values not larger than α divided by the total number of tests, i.e., the method rejects hypotheses having p-value $\leq \alpha/M$. This method has strong FWER control which means that it controls the FWER for any combination of true and false null hypotheses. Also, Bonferroni provides FWER control under any dependence structure of the p-values. The method is conservative, i.e., the probability of type I error is usually below the nominal α level.

Bonferroni is used in GWAS analysis because it is assumed that analyzing the whole genome is like performing only 10^6 tests (instead of the very much higher number of SNPs) and under such assumption, the FWER can be controlled by a Bonferroni adjustment with a genome-wide error level $\alpha = 5 \times 10^{-8}$. This assumption is based on the strong local correlations in the genome but the number of tests shouldn't be taken as universal since it depends on the correlation structure of the p-values and the kind of variant analysis we perform onto the genome (Hoggart et al., 2008; Lin, 2019; Sham and Purcell, 2014).

Bonferroni correction can be performed, using the R software, by the p.adjust function of the package stats (R Development Core Team, 2019) or by the package multcomp (Hothorn et al., 2008) and outside the R environment, by the software Myriads in its version v1.2 (Carvajal -Rodríguez, 2018).

2.1.3 Holm (Sequential Bonferroni)

This method is a sequential version of the Bonferroni. It is a step-down procedure, i.e., iterates beginning from the highest test value (smallest p-value). Like Bonferroni, Holm's method has strong FWER control at level α under any dependence structure of the p-values. It is more powerful than Bonferroni and should be in general recommended instead of it.

Regarding the application of the method in genomics, as a FWER control method is specially relevant for confirmatory (non-exploratory) experiments (Fig. 1). However, Holm's is very conservative and when the dependence structure is adequate (see below), more powerful variants should be used (Goeman and Solari, 2014).

Holm's sequential Bonferroni correction can be performed using the R software by the p.adjust function of the package stats (R Development Core Team, 2019) or by the package multcomp (Hothorn et al., 2008) and outside the R environment, by any of the versions of the software Myriads (Carvajal-Rodríguez, 2018).

2.1.4 Hommel

The Hommel method is a step-up procedure, i.e., iterates beginning from the lowest test value (largest p-value). Hommel's method has strong FWER control at level α, it operates under the assumption of independence of p-values or even positive dependence (through stochastic ordering, see Section 4).

Regarding the application of the method in genomics, it has the advantage of being more powerful than the Bonferroni and Holm methods but with the requirement of independence or at least positive dependence in the p-values.

Hommel's correction can be performed using the R software, by the p.adjust function of the package stats (R Development Core Team, 2019) or by the package Hommel (Goeman et al., 2019a) and outside the R environment, by the software Myriads v1.2 (Carvajal-Rodríguez, 2018).

2.1.5 MaxT

MaxT (Westfall and Young, 1993) is a permutation-based method that provides strong FWER control. MaxT does not impose any assumption on the dependence of the data and even the

assumption of uniformly distributed *p*-values is not required. Besides, MaxT is more powerful than the Holm and Hommel FWER-control methods. A limitation of the method is that it requires an invariance condition (interchangeability) that means that the distribution of the permuted data set should be identical to the original one. Thus, we cannot always define adequate (holding the invariance condition) permutations for all hypotheses and models. In general, only relatively simple experimental designs allow permutation tests to be used (Goeman and Solari, 2014). Another limitation is the computational cost when the number of tests is large. Thus, a key issue for the computational feasibility of MaxT is the number of permutations required to obtain acceptable accuracy (see Section 4 for details).

MaxT guarantees FWER control while providing more power than other FWER methods, so it is a recommended method for confirmatory (non-exploratory) genomics data analysis when appropriate permutations are available (Fig. 1).

MaxT correction can be performed, using the R software, by the mt.maxT function of the package multtest (Pollard et al., 2005) and outside the R environment, by the software Myriads v1.2 (Carvajal-Rodríguez, 2018).

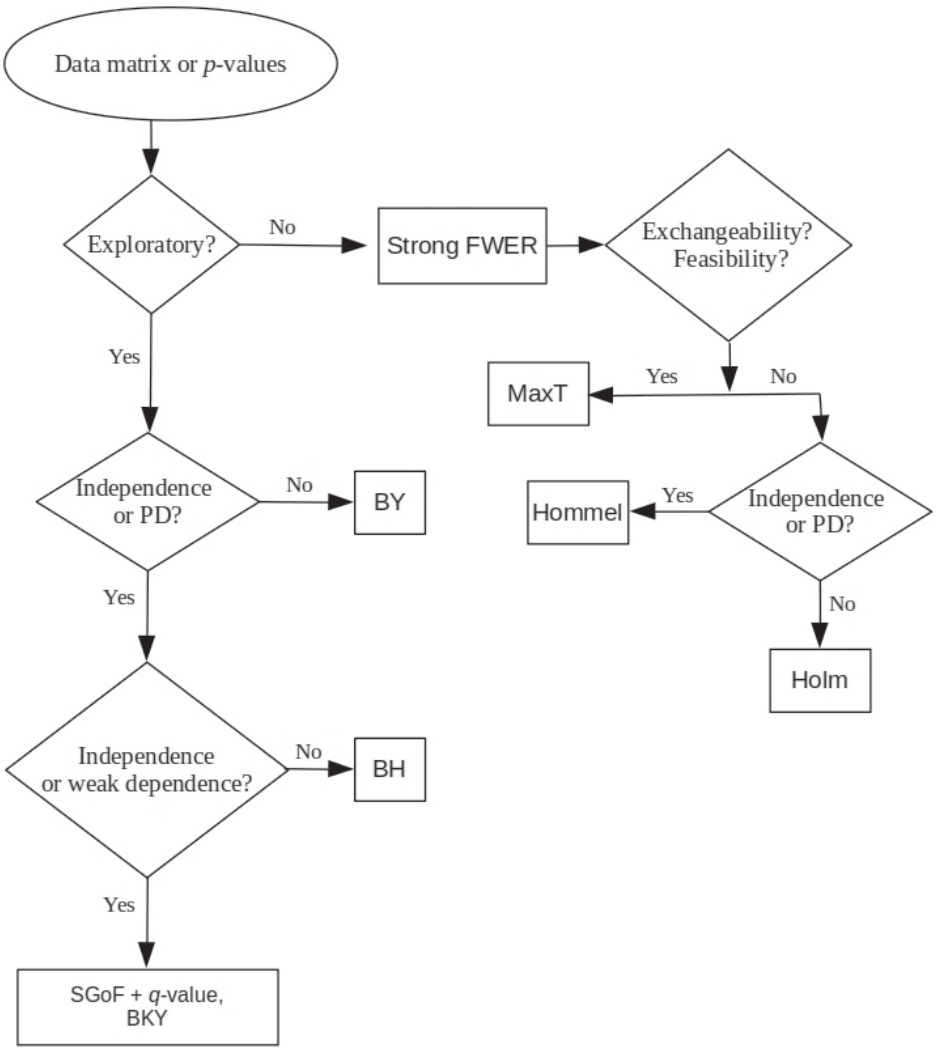

Figure 1 Decision flowchart for multiple testing procedures. PD: positive dependence. (Figure Courtesy, Myriads Manual, Carvajal-Rodríguez, 2018)

2.1.6 SGoF

The Sequential Goodness of Fit (SGoF) method is a step-down FWER-controlling procedure in the weak sense, which means that it has a strict control only under the complete null hypothesis, i.e., when all null hypotheses are true. Unlike methods with strong FWER control, SGoF increases its power with the number of tests and it can detect weak effects (Carvajal-Rodríguez et al., 2009; Carvajal-Rodríguez and de Uña-Álvarez, 2011; de Uña-Álvarez, 2011). However, the method requires the assumption of independence of p-values and can be too liberal when large blocks of correlated values are present (Carvajal-Rodríguez, 2018).

SGoF's weak FWER control may be useful if large power is preferred over type I error strong control. This scenario may occur when screening a large number of factors for detecting some important association that can be validated later; or simply when the researcher is primarily interested in an exploratory phase. Of course, when the requirements are confirmatory, only methods that keep strong control should be used (Fig. 1).

SGoF correction can be performed, using the R software, by the SGoF package (Castro-Conde and de Uña-Álvarez, 2014) and outside the R environment, by any of the versions of the software Myriads (Carvajal-Rodríguez, 2018).

2.2 Methods that Control False Discovery Rate (FDR)

In general, FDR control is less conservative and more powerful than FWER, specially when there are many false hypotheses, which makes FDR control a suitable tool for exploratory genomics studies where researchers are interested in selecting sets of promising hypotheses. From herein the FDR control methods are assumed to have strong control unless indicated otherwise. However, it is worth noting a key difference between FWER and FDR controlling procedures that is often omitted. If FWER is controlled at level α for a set of hypotheses then it is also controlled at the same level for any subset, including single hypotheses. This subset property is not true for FDR, which means that for a set with FDR controlled at level α it does not imply that a given subset is controlled at the same level (Finner and Roters, 2001). This is important because it means that the FWER for a set is also the FWER for each hypothesis in the set which is not true in the case of FDR. Thus, the FDR-adjusted p-value is a property of the rejected set not of any individual hypothesis (Goeman and Solari, 2014). Therefore, if statements on individual hypothesis are required as in confirmatory (non-exploratory) experiments, the FWER controlling methods should be preferable (Fig. 1).

Table 4 FDR control methods with different dependence assumptions and usability in terms of the exploratory or non-exploratory nature of the experiments

Method	*Dependence assumptions*	*Usability level*	*Software*
BH	Positive dependence	exploratory	p.adjust, Myriads
BY	None	exploratory	p.adjust, Myriads v1.2
BKY	Positive dependence	exploratory	cp4p, Myriads v1.2

2.2.1 Benjamini & Hochberg (BH)

The Benjamini & Hochberg (BH) method (Benjamini and Hochberg, 1995) is a step-up FDR controlling procedure under the assumption of independence of p-values or even positive dependence. BH is more powerful than Hommel.

BH correction can be performed, using the R software, by the p.adjust function of the package stats (R Development Core Team, 2019) and outside the R environment, by any of the versions of the software Myriads (Carvajal-Rodríguez, 2018).

2.2.2 Benjamini & Yekutieli (BY)

The Benjamini & Yekutieli (BY) method (Benjamini and Yekutieli, 2001) is a step-up FDR controlling procedure that is valid under any dependence structure. It is more conservative and has less power than the BH procedure, so BY should only be preferred when the dependence structure cannot be guaranteed (Fig. 1).

BY correction can be performed, using the R software, by the p.adjust function of the package stats, (R Development Core Team, 2019) and outside the R environment, by the software Myriads v1.2 (Carvajal-Rodríguez, 2018).

2.2.3 BKY adaptive two-stage method

The BH method controls FDR at level $\pi_0\alpha$, where $\pi_0 = M_0/M$ is the proportion of true nulls. The adaptive BH-like procedures intend to gain power over BH by estimating π_0 in a first step, and controlling FDR over $\pi_0 M$, instead of over M, in the second step. Benjamini, Krieger and Yekutieli (Benjamini et al., 2006) developed one of such procedures (BKY, see Section 4 for details). The BKY method is an adaptive BH-like procedure that controls FDR under independence or positive dependence and even remains conservative with relative good power when the degree of dependence is unknown (Blanchard and Roquain, 2009; Kim and van de Wiel, 2008).

The BKY correction can be performed, using the R software, by the adjust.p function (with pi0.method = "bky" argument) of the package cp4p (Gianetto et al., 2019) and outside the R environment, by the software Myriads v1.2 (Carvajal-Rodríguez, 2018).

2.3 Methods that Estimate the False Discovery Proportion (FDP)

FDR is the expectation of the false discovery proportion (FDP) unconditional on the occurrence of rejections. Under an FDR controlling procedure at level α, if for a subset of hypotheses the expected FDP is less or equal to α, we reject such hypotheses. For the FDP estimation, the procedure is reversed, it starts with the set of hypotheses candidate for rejection and finds an estimate Q for the FDP of that set. It can be argued that knowledge about the FDP is more relevant because it is directly related to the current experiment. That is, if we use any of the procedures of the previous section for controlling FDR at level α, we may obtain a set of rejected hypotheses. In this rejected set, the FDP is in average bounded by α, and the word "average" is key here, meaning that the FDR procedure is only indirectly (in average) describing the error rate of the rejected data. However, if we obtain an estimate of the FDP for a given set of hypotheses, it is a direct description of the error rate committed for such rejections (see FDP vs FDR in Section 4).

2.3.1 Storey's bayesian pFDR-estimation (*q*-values)

From the point of view of a researcher it seems desirable that when all nulls are true ($\pi_0 = 1$), the false discovery rate should be 1 because in this scenario any rejection is obviously a false one. However, the FDR when $\pi_0 = 1$ can be less than one. Even more, the researcher may not be interested in cases where there is no significant test. To solve this requirement the pFDR is defined as the expected FDP when at least one discovery has occurred (Storey, 2003). Defined in this way, and because we cannot control when the discoveries happened or not, the pFDR cannot be controlled a priori, contrary to the FDR in the BH and other FDR controlling procedures. Therefore, the computation of pFDR can be viewed as an FDP estimation (Fan et al., 2012).

The pFDR assumes independency, although seems to work well under weak dependence may have great variance under realistic dependence values (Goeman and Solari, 2014).

For estimating π_0, Storey's method uses a threshold parameter λ so the pFDR estimate has sometimes being called Storey-λ (Blanchard and Roquain, 2009).

Because the pFDR is not controlled a priori, adjusted *p*-values cannot be formally defined but the so-called *q*-values, that give an error measure for each statistic with respect to pFDR (see Section 4). Like the FDR controlling procedures, the Storey-λ FDP estimation lacks the subset property, i.e., the *q*-value is a property of the hypothesis in a given rejection set not of the individual hypotheses.

The Storey-λ *q*-values can be computed, using the R software, by the package *q*-value (Storey et al., 2020) and outside the R environment, by any of the versions of the software Myriads (Carvajal-Rodríguez, 2018).

2.3.2 Significance analysis of microarrays (SAM)

SAM (Tusher et al., 2001) is a method that estimates FDP by a permutation-based variant of the Storey-λ method with $\lambda = 0.5$ (Storey and Tibshirani, 2003a). Because it is permutation-based it adapts to the dependence structure. However, it is a known concern that the estimate of π_0 is less accurate under dependence (Schwartzman and Lin, 2011). See Section 4 for more details. SAM estimates can be computed, using the R software, by the package samr (Tibshirani et al., 2018).

2.3.3 Efrom's bayesian local FDR-estimation

The concept of local-FDR (Efron et al., 2001) solves the problem of lacking the subset property by providing estimates of FDP for individual hypotheses. The approach of Efrom has the additional requirement of assuming that the set of tests follows a mixture distribution of test statistics for the true and false hypotheses with prior probabilities π_0 and $1 - \pi_0$ respectively. Similar to Storey-λ, the local-FDR method suffers high variance in π_0 and in the FDP estimates distribution when the statistics are correlated, in such cases the empirical null distribution estimation should be preferred (Efron, 2007). Local-FDR estimates can be computed, using the R software, by the packages locfdr (Efron et al., 2015) and samr (Tibshirani et al., 2018).

Besides the above point estimates of FDP, there has been recent work on providing confidence intervals for the FDP (Goeman and Solari, 2011; Hemerik et al., 2019; Hemerik and Goeman, 2018). See Sections 4 and 5 for more details.

3 DEFINITIONS

We reproduce some relevant definitions with commentaries as they appear in the Myriads manual (Carvajal-Rodríguez, 2018).

3.1 Test of Hypotheses

A rule for deciding whether to accept or reject a hypothesis. The hypothesis under test is called the main or null hypothesis (Perezgonzalez, 2014; Sokal and Rohlf, 1981). The boundary for deciding between hypotheses is called the critical value of the test. The type I error expected using this critical value is called nominal type I error rate, or simply α level (see below).

3.2 Type I Error

The rejection of a true null hypothesis.

3.3 Nominal Type I Error Rate α

The user-supplied upper-bound for the type I error rate (Greenland, 2019).

3.4 Type II Error

The acceptance of a false null hypothesis.

3.5 Power of a Test

For detecting a true alternative hypothesis is the probability that the test will reject the null hypothesis. In the context of binary classification, e.g., medical testing, it is also called sensitivity or true positive rate. The power is one minus the probability of a type II error.

3.6 Observed *p*-value

Is the probability, under the null hypothesis H_0, that a test statistic would be equal to or more extreme than the observed value. This is equivalent to say that the observed p-value is the smallest nominal type I error level of the single hypothesis testing procedure that would allow rejection of H_0 given the test statistic value (Dudoit and Laan, 2008). A more formal and rigorous definition follows. Given a statistical model A and a tested hypothesis H_0, the observed *p-value* is the probability that the test statistic be equal or larger than its observed value in the current sample realization if every model assumption were correct, including H_0 (Greenland et al., 2016).

3.7 Complete Null Hypothesis

When all null hypotheses are true.

3.8 Control in the Weak Sense

Control in the weak sense occurs when the type I error rate is controlled at the specified level only under the complete null hypothesis.

3.9 Control in the Strong Sense

Control in the strong sense occurs when the type I error rate is controlled at the specified level for any combination of true and false null hypotheses.

3.10 Per-Family Error Rate (PFER)

The expected number of type I errors in the family of tests, i.e., PFER = $E(V)$.

3.11 Per-Comparison Error Rate (PCER)

The expected value of the number of type I errors divided by the number of tests, i.e., PCER = $E(V)/M$. If we are performing each test at level α, the PCER is equal or less than α. This is the expected type I error rate, also called EER in (Finner and Roters, 2001). In a multiple testing context, controlling the PCER is like not considering the multiple test setting at all, and consequently is more liberal (anti-conservative) than controlling FWER or FDR (see definitions below).

3.12 Family-Wise Error Rate (FWER)

Is the probability of at least one type I error in the set of tests, i.e., FWER = Pr $(V \geq 1)$. Controlling the FWER is a conservative strategy which means that the probability of rejecting the null hypotheses is below the nominal α level. There is a generalization of the FWER concept called k-FWER which is the probability of at least k false rejections. Obviously, FWER corresponds to k-FWER with $k = 1$ (Lehmann and Romano, 2005).

3.13 False Discovery Proportion (FDP)

FDP is the (unobserved) proportion of false rejections V among total rejections R, FDP = V/R, with the convention of FDP = 0 when $R = 0$.

3.14 False Discovery Rate (FDR)

FDR is the FDP averaged over all possible experimental replicates. More technically, it is the expected FDP unconditional on the occurrence of rejections. Because FDP equals 0 when $R = 0$, FDR can be expressed as

$$\text{FDR} = E\left[\frac{V}{R}\middle| R > 0\right] \cdot \Pr(R > 0) \tag{1}$$

as defined in (Benjamini and Hochberg, 1995) but see also (Storey, 2003). We will make explicit some differences between FDR control versus FDP estimation in Section 4.5.

3.15 Positive False Discovery Rate (pFDR)

The expected proportion of false rejections V among the rejections R conditioned on and least one rejection

$$\text{pFDR} = E\left[\frac{V}{R}\middle| R > 0\right] \tag{2}$$

Note that, because we ignore if any rejections will occur, pFDR cannot be controlled at any given threshold (Storey, 2002; 2003). The pFDR and FDR coincide when the number of tests is large enough to guarantee at least one rejection, i.e., $\Pr(R > 0) = 1$. However, under finite sample size $R = 0$ may occur, so in the estimation of pFDR when $R = 0$, the convention of substituting R by 1 is applied (Storey, 2002). Under the complete null, the value of pFDR is 1, provided that some rejection occur. The computation of pFDR can be considered as a point estimate of the FDP (Fan et al., 2012; Goeman and Solari, 2014).

3.16 Subset Property

A FWER-controlling multiple testing procedure (MTP) is said that has the subset property if, when rejecting a set of hypotheses, the FWER control is also guaranteed for any subset including the single hypotheses. An FDR controlling procedure has not necessarily the subset property but if it has, such procedure also controls the FWER at level α. That is, any MTP with the subset property is a FWER controlling procedure (Finner and Roters, 2001; Goeman and Solari, 2014).

4 TECHNICAL DETAILS AND ALGORITHMS

4.1 Assumptions of Multiple Testing Correction Methods

4.1.1 Uniformity

Let us note the M_0 p-values corresponding to true nulls as q_1, ..., q_{m0}, then these p-values should be uniformly distributed between 0 and 1, or stochastically greater than uniform if data are discrete, anyway they should satisfy

$$\Pr(q_i \leq t) \leq t \tag{3}$$

Most MTPs require the p-values satisfy (3). However, the uniformity may be only approximate specially for small sample sizes.

4.1.2 Dependence structure

Some MTPs as Bonferroni, Holm, BY, or permutation-based methods do not require any dependence assumption and work under independence or any form of general dependence of the p-values. These methods are in general less powerful. Methods as Hommel and BH, work under independence or some kind of positive dependence called positive dependence through stochastic ordering (PDS, aka positive regression dependence) on the subset of p-values of true null hypotheses (Benjamini and Yekutieli, 2001; Goeman and Solari, 2014; Sarkar, 2008). Methods as Storey-λ require weak positive dependence implying only local correlations (Schwartzman and Lin, 2011). Even methods working with some kind of positive dependence may fail when there are strong correlations in the p-values (Blanchard and Roquain, 2009).

4.2 Adjusted *p*-values and *q*-values

Adjusted p-values can be considered as the multiple testing analogous of the single hypothesis test observed p-values. The adjusted p-value $\tilde{p}_m$ of a hypothesis H_m is the smallest α (nominal type I error level measured as FWER or FDR) of the multiple hypotheses testing procedure at which one would reject the hypothesis, given the test statistic value (Dudoit and Laan, 2008; Tamhane and Gou, 2018). However, because of the subset property, the interpretation of the adjusted p-value is different for FWER and FDR controlling methods. In the FWER controlling methods, the adjusted p-value is a property of each single hypothesis, but in the FDR controlling methods the adjusted p-value is a property of the set of rejected hypotheses, not of each hypothesis. The concept of local FDR (Efron et al., 2001) solves this by providing estimates of FDP for individual hypotheses.

The q-value is the minimum pFDR that can occur when rejecting the statistic (Storey, 2002). Often, it is said that q-values are adjusted pFDR p-values, which is technically incorrect because pFDR cannot be controlled by a test procedure (Storey, 2003). Like the adjusted FDR p-values, the q-values are a property of the set of rejected hypotheses, not of each hypothesis.

4.3 FWER Methods

4.3.1 Bonferroni

Bonferroni's method consists of rejecting hypotheses only if they have raw p-value smaller than α/M. This method provides FWER control for the set of M hypotheses at α level under any dependence structure of the p-values. It conservatively controls FWER for any combination of true and false hypotheses, i.e., it controls the FWER at level $\pi_0\alpha$, where π_0 is the unknown

proportion of true null hypotheses in the set of tests. Therefore, if there are many false null hypotheses (π_0 is low) the method is very conservative. In general, Bonferroni will be very conservative if the *p*-values are positively correlated and less conservative if the *p*-values are independent or negatively correlated (Goeman and Solari, 2014). The adjusted *p*-value for a test *i* under Bonferroni is min (Mp_i, 1), where p_i is the raw *p*-value.

Algorithm with adjusted p-values

Compute the adjusted *p*-values $\tilde{p}_i = \min\{M \times p_i, 1\}$.

Reject hypotheses with adjusted *p*-value $\tilde{p}_i \leq \alpha$.

4.3.2 Holm

Holm's method works in step-down way by iterating the Bonferroni method as follows. Consider the sorted *p*-values $p_1 \leq p_2 \ldots \leq p_M$ and their corresponding sorted hypotheses. The hypothesis $i (i = 1, \ldots, M)$ will be rejected if its raw *p*-value is smaller than $\alpha/(M - i + 1)$, else the procedure ends.

Like Bonferroni, Holm's procedure has FWER control with the only assumption of uniformity of *p*-values [Eqn (3)]. Holm's method is less conservative and more powerful than Bonferroni and should be used instead (Goeman and Solari, 2014).

Algorithm with adjusted p-values

0. Set $i = 1$.
1. Sort the *p*-values $p_1 \leq p_2 \ldots \leq p_M$.
2. Compute the adjusted *p*-value $\tilde{p}_i = \min\{(M - i + 1) \times p_i, 1\}$.
3. Enforce monotonicity $\tilde{p}_i = \max\{\tilde{p}_h\}$, $h = 1, \ldots, i$.
4. If $i < M$ increase i and repeat from step 2.
5. End.

Reject hypotheses H_i with adjusted *p*-value $\tilde{p}_i \leq \alpha$.

4.3.3 Hommel

Hommel's method is a step-up procedure that has strong FWER control under the assumption of independence of *p*-values or even positive dependence (Goeman and Solari, 2014; Tamhane and Gou, 2018). The classical algorithm for computing Hommel adjusted *p*-values is the Wright algorithm (Wright, 1992) which is time consuming. However, a recent work by (Meijer et al., 2019) provides a linear time algorithm for computing the adjusted *p*-values. Here we present the Hommel step-up adjusted *p*-values as computed from (Wright, 1992).

Algorithm with adjusted p-values

0. Sort the *p*-values $p_1 \leq p_2 \ldots \leq p_M$
1. Initially set $\tilde{p}_i = p_i$ for all i.
2. For each $m = M, M - 1, \ldots, 2$ (in that order), do:
   - 2a. For each $i > (M - m)$ do:
     Compute $c_i = (mp_i)/(m + i - M)$
   - 2b. $c_{\min} = \min\{c_i\}$, $i = M - m + 1, \ldots, M$. // for this subset
   - 2c. For each $i > (M - m)$ do:
     If $\tilde{p}_i < c_{\min}$ then $\tilde{p}_i = c_{\min}$
   - 2d. For $i \leq (M - m)$ do:
     (i) $c_i = \min(c_{\min}, mp_i)$
     (ii) if $\tilde{p}_i < c_i$ then $\tilde{p}_i = c_i$.
3. End

Reject hypotheses H_i with adjusted *p*-value $\tilde{p}_i \leq \alpha$.

4.3.4 SGoF for weak FWER control

The SGoF procedure is a step-down FWER-control method in the weak sense, i.e., under the complete null hypothesis. It was originally developed in (Carvajal-Rodríguez et al., 2009) and different variants have been developed since then (Carvajal-Rodríguez and de Uña-Álvarez, 2011; Castro-Conde et al., 2017; Castro-Conde and de Uña-Álvarez, 2015a; de Uña-Álvarez, 2012). The adjusted p-values can be computed following the algorithm in (Castro-Conde and de Uña-Álvarez, 2015b). There is also an efficient algorithm that provides an upper-bound for the adjusted p-values which can be useful when the number of tests is very high (Carvajal-Rodríguez, 2018). The statistical properties of the SGoF procedure were described in (de Uña-Álvarez, 2011, 2012).

4.3.5 A note on permutation based multiple testing

Instead of making assumptions about the dependence we can adapt the procedure to the observed dependence structure by using a permutation test. However, it is required that the null invariance condition or exchangeability (Westfall and Troendle, 2008) is satisfied, i.e., shuffling the observations should keep the data just as likely as the original set under the null hypothesis. Also, the power of permutation testing comes with a computational cost for moderate sample sizes when all permutations are performed. Thus, an alternative are Monte Carlo permutation tests that sample a sufficient number of permutations instead of doing all. The number of permutations varies, depending on the kind of test, the response variable, and the sample size n. For example, for the case-control scenario we have $B = n!/(n_1!n_2!)$ permutations of the n_1 case and n_2 control labels. For more details on re-sampling algorithms, the following references may be consulted (Dudoit et al., 2003; Ge et al., 2003; Romano and Wolf, 2005).

4.3.6 MaxT

As we have seen in Section 2, MaxT (Westfall and Young, 1993) is a powerful permutation-based method that provides strong FWER control. It does not impose any assumption on the dependence of the data and even the assumption of uniformly distributed p-values is not required.

A permutation algorithm for step-down maxT adjusted p-values can be found in (Dudoit et al., 2003, Box 2 in Section 2.6) and a recent efficient algorithm for resampling-based step-down adjusted p-values can be found in Algorithm 4.1 in (Romano and Wolf, 2016).

So far, we have reviewed various methods that provide FWER control, there are also generalized step-down methods for controlling k-FWER, the reader is referred to (Romano and Wolf, 2007) for a review of these methods.

4.4 FDR Methods

4.4.1 Benjamini-Hochberg (B-H) FDR-control method

The (unconditional) FDR (Benjamini and Hochberg, 1995) is the expected false discovery proportion (FDP, see Definitions 3.13 and 3.14). It is unconditional because it is not conditioned on the existence of rejections (Zaykin et al., 2000). The Benjamini-Hochberg (BH) is a step-up procedure that controls FDR at a desired level α.

Algorithm with adjusted p-values

0. Set $i = M$.
1. Sort the p-values $p_1 \leq p_2 \ldots \leq p_M$.
2. Compute the adjusted p-value $\tilde{p}_i = \min\{(M/i) \times p_i, 1\}$
3. Enforce monotonicity $\tilde{p}_i = \min\{\tilde{p}_h\}$, $h = i, \ldots, M$.
4. If $i > 1$ decrease i and repeat from step 2.
5. End.

Reject hypotheses H_i with adjusted p-value $\tilde{p}_i \leq \alpha$.

Interestingly, the FDR can be written in terms of specificity and sensitivity (power) (Storey and Tibshirani, 2003b), so

$$\mathrm{FDR} = E\left[\frac{M_0 \cdot (1 - specificity)}{M_0 \cdot (1 - specificity) + M_1 \cdot sensitivity}\right] \tag{4}$$

where M_0 is the number of true null hypotheses and $M_1 = M - M_0$ the number of true alternative hypotheses.

Recalling that the false positive rate $\alpha = 1 -$ specificity, and that sensitivity is 1 minus the probability of type 2 error (β) then we have

$$\mathrm{FDR} = E\left[\frac{M_0\alpha}{M_0\alpha + M_1(1-\beta)}\right]$$

If we divide M_0 and M_1 by M the quotient does not change and so

$$\mathrm{FDR} = E\left[\frac{\pi_0\alpha}{\pi_0\alpha + \pi_1(1-\beta)}\right] \tag{5}$$

where $\pi_0 = M_0/M$ and $\pi_1 = M_1/M = 1 - \pi_0$.

For any given error level α and statistical power $1 - \beta$, FDP increases with π_0 and decreases with π_1. When $\pi_0 = 1$, FDP = 1 so, we can ask, how can FDR, which is the FDP expectation, be controlled at any level when $\pi_0 = 1$? The answer to this important question is given in subsection 4.5.

4.4.2 Benjamini and Yekutieli (BY)

The Benjamini and Yekutieli (BY) method (Benjamini and Yekutieli, 2001) is a step-up FDR controlling procedure that, unlike BH, is valid under any dependence structure. The price for this is that BY has less power than BH, so BY should only be preferred when either independence or positive dependence structure cannot be guaranteed (Fig. 1).

The adjusted p-values for BY can be computed with the same algorithm as for BH just changing step 2 to be

$$\tilde{p}_i = \min\left\{\left(\frac{kM}{i}\right) \times p_i, 1\right\} \text{ with } k = \sum_{j=1}^{M} \frac{1}{j} \quad .$$

4.4.3 BKY adaptive two-stage method

As seen in Section 2, the BKY method is an adaptive two-stage BH-like procedure that controls FDR under independence or positive dependence with relative good power and conservativeness even when the degree of dependence is unknown (Benjamini et al., 2006; Blanchard and Roquain, 2009; Kim and van de Wiel, 2008).

The adjusted p-values for BKY can be computed with the same algorithm as for BH in two different steps. First, the algorithm of BH is utilized at level $\alpha' = \alpha/(1 + \alpha)$. Let r_1 be the number of rejected hypotheses by $\mathrm{BH}_{\alpha'}$ then estimate π_0 as $p_0 = (M - r_1)/M$. If $p_0 = 1$ do not reject any hypothesis and finish; if $p_0 = 0$ rejects all hypotheses and finish; otherwise, perform a second BH at level $\alpha^* = \alpha'/p_0$. In doing so we can compute the adjusted p-values following the BH algorithm just changing the step 2 to be

$$\tilde{p}_i = \min\left\{\left(\frac{(1+\alpha)\, p_0 M}{i}\right) \times p_i, 1\right\}$$

Reject hypotheses H_i with adjusted p-value $\tilde{p}_i \leq \alpha$.

In addition to the BKY method and those reviewed in (Benjamini et al., 2006) there are various other kinds of adaptive procedures, the interested reader may consult (Blanchard and Roquain, 2009; Kang, 2020a).

4.4.4 Weighted FWER and FDR methods

Sometimes there is available meaningful prior information related to the hypotheses being tested so that not all null hypotheses have the same importance. Consequently, weighted multiple hypothesis testing procedures have being developed for FWER as well for FDR control (Genovese et al., 2006; Kang et al., 2009). When the weights are informative these methods are usually more powerful than their unweighted counterparts.

There are in general two strategies for estimating the weights: external weights, where prior information (based on scientific knowledge or prior data) exists for specific hypotheses; and estimated weights, where some informative covariates are used to construct weights (Ignatiadis et al., 2016; Ignatiadis and Huber, 2017; Korthauer et al., 2019; Roeder and Wasserman, 2009).

Usually, the weights must satisfy

$$\frac{1}{M}\sum_{j=1}^{M} w_j = 1 \quad \text{with } w_j \geq 0$$

and the weighted p-values are defined as $\frac{p_i}{w_i}$. Adjusted weighted p-values can be computed for different FWER and FDR procedures, see for example (Genovese et al., 2006; Kang et al., 2009).

4.5 FDP vs FDR

Consider the complete null hypothesis where all nulls are true ($\pi_0 = 1$) or equivalently, all rejections are false, $V = R$ and so FDP = $V/R = 1$. In the previous subsection when talking about BH we have asked, how can FDR, which is the FDP expectation, be controlled at any level when $\pi_0 = 1$? To put it clear, if FDP is 1, how can we assure FDR ≤ 0.05?

First, it is important to recall that FDR is unconditional on the number of rejections (Zaykin et al., 2000). In other words, FDR can be defined as an expectation independently of the observed tests, with or without rejections. It maintains the control because it is an average, so, if we are controlling the FDR at level α we expect that, on an average, the proportion of false discoveries is upper-bounded by α. This average, allow me to insist, is what we call FDR, and is the average what is maintained below α. In the case of the BH procedure, the FDR upper bound is $\pi_0\alpha$ (Benjamini et al., 2006).

Let's see an example, consider performing an *in silico* experiment or simulation, for which all null hypotheses are true, i.e., the complete null hypothesis ($\pi_0 = 1$). This experiment consists in $M = 100$ independent tests each performed at $\alpha = 0.05$. If we do not apply an MTP, we expect 5 false rejections in 100 tests (PFER = 5). Under the complete null the FDR = FWER $= 1 - (1 - 0.05)^{100} = 0.99$. Furthermore, we repeated the same kind of experiment 1000 times. On each occasion the FWER and FDR is 0.99.

Because of such a high error we decide to control the FDR at 0.05 level applying the BH procedure. However, under the complete null we should be aware that in any experiment in which we have some rejection ($R > 0$) the observed proportion of false discoveries is FDP = 1. On the contrary, in the experiments without rejections ($R = 0$), the FDR takes value 0 by definition and so does the FDP. Thus, applying $BH_{0.05}$ we expect that in 95% of our experiments there are no rejection so FDP = 0 while 5% have at least one rejection so FDP = 1. Then we can compute the expected FDP which is FDR = E(FDP) = $0.95 \times 0 + 0.05 \times 1 = 0.05$. This is how the FDR can be controlled at level α when all nulls are true or, in general, at level $\alpha\pi_0$, whatever the

combination of nulls and false hypotheses. Under the complete null, at each experiment the observed FDP is 0 or 1, although the FDR is being controlled as an expectation of 0.05.

Let us consider now what happens with pFDR, which is a conditional FDR. The pFDR estimates the FDP for a given rejected set, which means that in our *in silico* lab we will obtain an FDP value for each performed experiment having rejections (Carvajal-Rodríguez and de Uña-Alvarez, 2011; Schwartzman, 2012; Storey, 2003; Zaykin et al., 2000). Under the complete null, the probability α of false rejection, matches the probability of rejection, $\Pr(R > 0) = \alpha$, and recalling the definition of FDR in terms of pFDR, i.e., FDR = pFDR × $\Pr(R > 0)$ and solving for pFDR, we get pFDR = FDR/$\Pr(R > 0) = \alpha/\alpha = 1$. It is clear that, given at least one rejection in our experiment, although the BH controls the FDR at level α, we are still committing a pFDR = 1, independently of α and the statistical power.

Whether FDR or pFDR are the quantities of interest has been disputed, e.g., Storey argues that the interest relies in situations where there is at least one rejection. On the other side, (Dudoit et al., 2008) argues that pFDR being equal to one under the complete null hypothesis impedes its control under this testing scenario. Besides, the FDR reduces to the FWER = $\Pr(V > 0)$ under the complete null.

4.6 FDP Estimation Methods

4.6.1 Storey's Bayesian pFDR-estimation (*q*-values)

The pFDR was defined as the expected FDP when at least one discovery has occurred (Storey, 2003; Storey and Tibshirani, 2001). The pFDR can be estimated as $\text{pFDR}(\alpha) = p_0 tM/R(t)$, where p_0 is an estimate of π_0 and $R(t)$ is the total number of rejections under probability threshold t. Thus, we only need to estimate π_0. It is worth noting that under certain general conditions, the pFDR can be expressed as a Bayesian posterior probability, where π_0 is the prior for the posterior probability of rejecting a true null (Storey, 2003). We have mentioned in Section 2 that Storey uses a threshold parameter λ for estimating π_0. The value of the parameter λ can be fixed or automatically computed by bootstrap or fitting a cubic spline (Storey and Tibshirani, 2003b; Storey, 2002). Besides λ-estimation, several other methods has been proposed for estimating π_0, some of them have been reviewed in (Carvajal-Rodríguez, 2018; Carvajal-Rodríguez and de Uña-Álvarez, 2011; Friguet and Causeur, 2011; Kang, 2020b). The λ-estimation and related methods assume independence or weak dependence implying local correlations. Under more realistic dependence, the FDP estimation can have very large variance, skewness and bias (Owen, 2005).

Analogous to the FDR adjusted *p*-values, the *q*-value is the minimum pFDR that can occur when the given statistic is rejected for the set of rejection regions (Storey, 2002). As already commented, the *q*-value, although interesting, is a property of the hypothesis within the specific rejected set not of the individual hypothesis itself.

4.6.2 SAM

The significance analysis of microarrays (SAM) method (Tusher et al., 2001) does not control neither estimate FDR but the expected number of false positives PFER = $E(V)$ (Dudoit et al., 2003). This value $E(V)$ divided by the number of rejections R is an estimate of the FDP although its reliability is not clear under general dependence conditions (Blanchard and Roquain, 2009; Goeman and Solari, 2014; Kim and van de Wiel, 2008). The SAM algorithm is given in (Dudoit et al., 2003). The procedure has been recently extended by the addition of upper bounds to the FDP estimation (Hemerik and Goeman, 2018). There is also an extended SAM software R package called confSAM.

4.6.3 Efrom's Bayesian local FDR-estimation (empirical null estimation)

The Benjamini and Hochberg's (1995) false discovery rate (FDR) is based on tail area properties (tail area FDR) while local false discovery rates are FDR based on densities. Given the test statistic t, the model of a mixture density is defined for the unaffected (null) and for the affected treatment of interest

$$f(z) = \pi_0 f_0(t) + \pi_1 f_1(t).$$

The model requires an estimate of the "null density" $f_0(t)$ which is done by empirical null estimation, for example, microarray data structures allow to estimate the density by permutation. The main advantage of the empirical null estimation is the robustness to dependence. The value of the proportion of true nulls π_0 of false hypotheses are estimated from data. However, like previous estimates both the density and π_0 may be highly variable when the p-values are correlated (Efron, 2005).

4.7 Relaxing the Continuity Assumption: Discrete *p*-values

Classic MTPs, including BH, were developed under the assumption that the p-values are uniformly and continuously distributed when the null is true. However, examples of discrete test statistics are common in genomics and related biomedical sciences (He and Heyse, 2019; Liang, 2016). Methods developed under the assumption of continuity could be too conservative when the p-values are discrete, and may not be as powerful as one would hope for. Several authors have demonstrated the advantages of using discrete information properties when the data is highly discrete, see for example (Westfall and Troendle, 2008; Westfall and Wolfinger, 1997). Discrete version of different MTPs has been developed both for FWER (Castro-Conde et al., 2017; He and Heyse, 2019; Zhu and Guo, 2020) and for FDR control and estimation (Chen, 2020; Chen and Sarkar, 2020; Döhler, 2018; Liang, 2016).

5 RECENT DEVELOPMENTS

In recent years, several multiple testing improvements indicate the strength of the field within the big biological data context. Regarding computational efficiency, new algorithms were recently proposed, e.g., for the Hommel's procedure in linear time (Meijer et al., 2019) or for resampling-based step-down adjusted p-values (Romano and Wolf, 2016), or the generalized FWER error control (k-FWER) and false positives and negatives control (Song and Fellouris, 2019). Regarding FDR, some new weighted covariate methods as IHW (Ignatiadis et al., 2016) that reduces to the BH procedure when the covariate is completely uninformative; similarly, BL (Boca and Leek, 2018) reduces to the Store's q-value; also, the functional FDR incorporates informative variables when available, for computing FDR and q-values (Chen et al., 2019). Some of these new FDR methods has been recently benchmarked in (Korthauer et al., 2019).

The estimation of FDP under strong dependence has been recently studied (Fan et al., 2019, 2012; Fan and Han, 2017). Another avenue of research come from the adaptive control of FWER and FDR by assuming dependency structure by blocks (Guo and Sarkar, 2019), or by estimation of confidence bounds for the false discovery proportion (Goeman et al., 2019b; Goeman and Solari, 2011; Hemerik and Goeman, 2018). Also, when there are logical relations among the hypotheses, the control can be exerted by hierarchically ordering the hypotheses and by graphical approaches (Tamhane and Gou, 2018). The different types of prior information, e.g., weights, dependence structure, proportion of nulls, hypotheses subgroups, can be combined for FDR control of grouped hypotheses (Ramdas et al., 2019).

Within the field of integrative genomics, the simultaneous analysis of multiple data sets boosted the extension of MTPs for multivariate *p*-values (Chi, 2008; Phillips and Ghosh, 2014; Richardson et al., 2016; Rudra et al., 2019; Xia et al., 2019). Last but not least, multiple hypotheses testing can also be done using the Bayesian counterpart of *p*-values, the Bayes factors so called *e*-values, and different procedures have been recently proposed under this setting (Vovk and Wang, 2019a,b).

6 CONCLUSIONS

We have reviewed classical and some of the most important methods for controlling type I error in multiple hypotheses testing. Though there is plenty of different methods, the main avenue for controlling type I error goes through FWER or FDR control and/or false discovery proportion (FDP) estimation. Meanwhile, we provided a map for deciding the best strategy depending on the research interest, either exploratory or confirmatory, the assumptions that can be made, the availability of prior information and the kind of statistical test. We either explicitly gave or referenced, some well-known algorithms for computing the adjusted *p*-values.

In addition, we made explicit some key differences between methods often omitted in the literature. First, FDR control may be less conservative than FWER which makes FDR control a suitable tool for exploratory genomics studies where we are interested in selecting sets of promising hypotheses; even more, if independence or weak dependence is guaranteed, the SGoF method can be a valuable and powerful exploratory alternative. On the contrary, FWER strong control is specially relevant for confirmatory (non-exploratory) experiments. Obviously, when possible, the most powerful and flexible FWER controlling variant, MaxT, should be used; alternatively, if permutation techniques are not allowable but the dependence structure is adequate, the Hommel's procedure is a good option also.

Second, the subset property states that, if FWER is controlled at level α for a set of hypotheses then it is also controlled at the same level for any subset. This means that we can associate the FWER control to a single hypothesis so that by rejecting hypothesis *i*, the type I error is bounded by α. The same subset property is not true for FDR, so that for a set with FDR controlled at level α, it does not mean that a given subset, e.g., a single hypothesis, is controlled at the same level. While the local FDR point estimate variants have the subset property, they are less powerful and more challenging to estimate accurately.

Third, recall that FDR is the FDP expectation and only controls the false positive rate in average which implies that the actual proportion of false discoveries in the rejected set can be substantially larger than the desired level, especially if the proportion of true nulls is large, e.g., under the complete null, the percentage of false discoveries may be 100% under an FDR control of 5%. The pFDR and *q*-value improves power over FDR procedures (as BH and BY) because, like adaptive methods, estimate the proportion of true nulls π_0 to control false positives from this proportion. Because pFDR is conditioned on occurrence of rejections it is more suited to error estimation than to error control; there are several techniques for FDP estimation providing an alternative strategy for managing type I error.

Finally, multiple testing is a very active research field, it seems that multiple testing procedures and *p*-value (and *e*-value) adjustment are here to stay. They are already important contributors for enhancing the reproducibility and reliability of scientific research.

Acknowledgements

I wish to thank Sonia Prado and Pili Alvariño for their comments on the manuscript. This research was supported by the Ministerio de Economía y Competitividad (CGL2016-75482-P)

and Xunta de Galicia (Grupo de Referencia Competitiva, ED431C 2020/05, Centro Singular de Investigación de Galicia accreditation 2019-2022), and by the European Union (European Regional Development Fund - ERDF, "Unha maneira de facer Europa")

REFERENCES

Benjamini, Y. and Y. Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological). 57: 289–300.

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

Benjamini, Y., A. Krieger and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. Biometrika. 93: 491–507.

Blanchard, G. and E. Roquain (2009). Adaptive False Discovery Rate Control under Independence and Dependence. Journal of Machine Learning Research. 10: 2837–2871.

Boca, S.M. and J.T. Leek (2018). A direct approach to estimating false discovery rates conditional on covariates. PeerJ. 6: e6035. https://doi.org/10.7717/peerj.6035.

Carvajal-Rodríguez, A., J. de Uña-Álvarez and E. Rolan-Alvarez (2009). A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests. BMC Bioinformatics. 10: 209.

Carvajal-Rodríguez, A. and J. de Uña-Álvarez (2011). Assessing significance in high-throughput experiments by sequential goodness of fit and q-value estimation. PLoS One. 6: e24700.

Carvajal-Rodríguez, A., (2018). Myriads: p-value-based multiple testing correction. Bioinformatics. 34: 1043–1045.

Castro-Conde, I. and J. de Uña-Álvarez (2014). sgof: An R Package for Multiple Testing Problems. The R Journal. 6: 96–113.

Castro-Conde, I. and J. de Uña-Álvarez (2015a). Power, FDR and conservativeness of BB-SGoF method. Computational Statistics. 30: 1143–1161.

Castro-Conde, I. and J. de Uña-Álvarez (2015b). Adjusted p-values for SGoF multiple test procedure. Biometrical Journal. 57: 108–122.

Castro-Conde, I., S. Döhler and J. de Uña-Álvarez (2017). An extended sequential goodness-of-fit multiple testing method for discrete data. Stat Methods Med Res. 26: 2356–2375.

Chen, X., D.G. Robinson and J.D. Storey (2019). The functional false discovery rate with applications to genomics. Biostatistics.

Chen, X. (2020). False discovery rate control for multiple testing based on discrete *p*-values. Biom. J. 62(4): 1060–1079.

Chen, X. and S.K. Sarkar (2020). On Benjamini–Hochberg procedure applied to mid p-values. Journal of Statistical Planning and Inference. 205: 34–45.

Chi, Z., (2008). False discovery rate control with multivariate p-values. Electron. J. Statist. 2: 368–411.

de Uña-Álvarez, J. (2011). On the Statistical Properties of SGoF Multitesting Method. Stat. Appl. Genet. Mol. Biol. 10(1): 18.

de Uña-Álvarez, J. (2012). The Beta-Binomial SGoF method for multiple dependent tests. Stat. Appl. Genet. Mol. Biol. 11: 1–32.

Döhler, S., (2018). A discrete modification of the Benjamini–Yekutieli procedure. Econometrics and Statistics. 5: 137–147.

Dudoit, S., J.P. Shaffer and J.C. Boldrick (2003). Multiple Hypothesis Testing in Microarray Experiments. Statist. Sci. 18: 71–103.

Dudoit, S., H.N. Gilbert and M.J. van der Laan (2008). Resampling-based empirical bayes multiple testing procedures for controlling generalized tail probability and expected value error rates: Focus on the false discovery rate and simulation study. Biom. J. 50: 716–744.

Dudoit, S. and M.J. van der Laan, (2008). Multiple Testing Procedures with Applications to Genomics. Springer Series in Statistics. Springer, New York.

Efron, B., R. Tibshirani, J.D. Storey and V. Tusher (2001). Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association. 96: 1151–1160.

Efron, B. (2005). Local false discovery rates.

Efron, B. (2007). Correlation and Large-Scale Simultaneous Significance Testing. Journal of the American Statistical Association. 102: 93–103.

Efron, B., B. Turnbull, B. Narasimhan and K. Strimmer (2015). locfdr: Computes Local False Discovery Rates.

Fan, J., X. Han, and W. Gu (2012). Estimating False Discovery Proportion Under Arbitrary Covariance Dependence. J. Am. Stat. Assoc. 107: 1019–1035.

Fan, J. and X. Han (2017). Estimation of the false discovery proportion with unknown dependence. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 79: 1143–1164.

Fan, J., Y. Ke, Q. Sun, and W.X. Zhou (2019). FarmTest: Factor-Adjusted Robust Multiple Testing with Approximate False Discovery Control. J. Am. Stat. Assoc. 114(528): 1880–1893.

Finner, H. and M. Roters (2001). On the False Discovery Rate and Expected Type I Errors. Biometrical Journal. 43: 985–1005.

Friguet, C. and D. Causeur (2011). Estimation of the proportion of true null hypotheses in highdimensional data under dependence. Computational Statistics & Data Analysis. 55: 2665– 2676.

Ge, Y., S. Dudoit and T.P. Speed (2003). Resampling-based multiple testing for microarray data hypothesis. Test. 12: 1–44.

Genovese, C.R., K. Roeder and L. Wasserman (2006). False Discovery Control with p-Value Weighting. Biometrika. 93: 509–524.

Gianetto, Q.G., F. Combes, C. Ramus, C. Bruley, Y. Couté and T. Burger (2019). cp4p: Calibration Plot for Proteomics.

Goeman, J.J. and A. Solari (2011). Multiple Testing for Exploratory Research. Statist. Sci. 26: 584–597.

Goeman, J.J. and A. Solari (2014). Multiple hypothesis testing in genomics. Statistics in Medicine. 33: 1946–1978.

Goeman, J., Meijer, Rosa, Krebs and Thijmen (2019a). hommel: Methods for Closed Testing with Simes Inequality, in Particular Hommel's Method. R package version, 1.

Goeman, J., R. Meijer, T. Krebs and A. Solari (2019b). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. Biometrika. 106: 841–856.

Greenland, S., S.J. Senn, K.J. Rothman, J.B. Carlin, C. Poole, S.N. Goodman, et al., (2016). Statistical tests, P-values, confidence intervals, and power: A guide to misinterpretations. European Journal of Epidemiology. 31: 337–350.

Greenland, S., (2019). Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values. The American Statistician. 73: 106–114.

Guo, W. and S. Sarkar (2019). Adaptive controls of FWER and FDR under block dependence. Journal of Statistical Planning and Inference.

He, L. and J.F. Heyse (2019). Improved power of familywise error rate procedures for discrete data under dependency. Biometrical Journal. 61: 101–114.

Hemerik, J. and J.J. Goeman (2018). False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 80: 137–155.

Hemerik, J., A. Solari and J.J. Goeman (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. Biometrika. 106: 635–649.

Hoggart, C.J., T.G. Clark, M.D. Iorio, J.C. Whittaker and D.J. Balding (2008). Genome-wide significance for dense SNP and resequencing data. Genetic Epidemiology. 32: 179–185.

Hothorn, T., F. Bretz and P. Westfall (2008). Simultaneous inference in general parametric models. Biometrical Journal. 50: 346–363.

Ignatiadis, N., B. Klaus, J. Zaugg and W. Huber (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat Methods. 13: 577–580.

Ignatiadis, N. and W. Huber (2017). Covariate powered cross-weighted multiple testing with false discovery rate control [WWW Document]. URL /paper/Covariate-powered-cross-weightedmultiple-testing-Ignatiadis-Huber/0021dcbefe3bdc7a00ea347894d26cb54ca187d8 (accessed 2.15.20).

Kang, G., K. Ye, N. Liu, D.B. Allison and G. Gao (2009). Weighted Multiple Hypothesis Testing Procedures. Stat Appl Genet Mol Biol. 8.

Kang, J., (2020a). Two-stage false discovery rate in microarray studies. Communications in Statistics-Theory and Methods. 49: 894–908.

Kang, J., (2020b). Comparison of methods for the proportion of true null hypotheses in microarray studie. Communications for Statistical Applications and Methods. 27: 141–148.

Kim, K.I. and M. van de Wiel (2008). Effects of dependence in high-dimensional multiple testing problems. BMC Bioinformatics. 9: 114.

Korthauer, K., P.K. Kimes, C. Duvallet, A. Reyes, A. Subramanian, M. Teng, et al. (2019). A practical guide to methods controlling false discoveries in computational biology. Genome Biology. 20: 118.

Larson, H.J., (1982). Introduction to Probability Theory and Statistical Inference, Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley.

Lehmann, E.L. and J.P. Romano (2005). Generalizations of the familywise error rate. Ann. Statist. 33: 1138–1154.

Liang, K. (2016). False discovery rate estimation for large-scale homogeneous discrete p-values. Biometrics. 72: 639–648.

Lin, D.Y., (2019). A simple and accurate method to determine genomewide significance for association tests in sequencing studies. Genetic Epidemiology. 43: 365–372.

MacDonald, P.W., K. Liang and A. Janssen (2019). Dynamic adaptive procedures that control the false discovery rate. Electronic Journal of Statistics. 13: 3009–3024.

Meijer, R.J., T.J.P. Krebs and J.J. Goeman (2019). Hommel's procedure in linear time. Biometrical Journal. 61: 73–82.

Owen, A.B., (2005). Variance of the Number of False Discoveries. Journal of the Royal Statistical Society. Series B (Statistical Methodology). 67: 411.

Perezgonzalez, J.D., (2014). A reconceptualization of significance testing. Theory & Psychology. 24: 852–859.

Phillips, D. and D. Ghosh (2014). Testing the disjunction hypothesis using Voronoi diagrams with applications to genetics. The Annals of Applied Statistics. 8: 801–823.

Pollard, K., S. Dudoit and M.J. van der Laan (2005). Multiple Testing Procedures: R multtest Package and Applications to Genomics, Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer.

R Development Core Team, (2019). R: A language and environment for statistical computing. Vienna, Austria.

Ramdas, A.K., R.F. Barber, M.J. Wainwright and M.I. Jordan (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. The Annals of Statistics. 47: 2790–2821.

Richardson, S., G.C. Tseng and W. Sun (2016). Statistical Methods in Integrative Genomics. Annual Review of Statistics and Its Application. 3: 181–209.

Roeder, K. and L. Wasserman (2009). Genome-Wide Significance Levels and Weighted Hypothesis Testing. Stat Sci. 24: 398–413.

Romano, J.P. and M. Wolf (2005). Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. Journal of the American Statistical Association. 100: 94–108.

Romano, J.P. and M. Wolf (2007). Control of generalized error rates in multiple testing. Ann. Statist. 35: 1378–1408.

Romano, J.P. and M. Wolf (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. Statistics & Probability Letters. 113: 38–40.

Rudra, P., E. Cruz-Cortés, X. Zhang and D. Ghosh (2019). Multiple testing approaches for hypotheses in integrative genomics. WIREs Computational Statistics n/a, e1493.

Sarkar, S.K., (2008). On Methods Controlling the False Discovery Rate. Sankhyā: The Indian Journal of Statistics, Series A (2008-). 70: 135–168.

Schwartzman, A. and X. Lin (2011). The effect of correlation in false discovery rate estimation. Biometrika. 98: 199–214.

Schwartzman, A., (2012). Comment: FDP vs FDR and the Effect of Conditioning. Journal of the American Statistical Association. 107: 1039–1041.

Sham, P.C. and S.M. Purcell (2014). Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 15: 335–346.

Sokal, R.R. and F.J. Rohlf (1981). Biometry, Second. ed. W. H. Freeman and Co., New York.

Song, Y. and G. Fellouris (2019). Sequential multiple testing with generalized error control: An asymptotic optimality theory. Ann. Statist. 47: 1776–1803.

Storey, J.D. and R. Tibshirani (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. Technical Report 2001–2028, Department of Statistics, Stanford University.

Storey, J.D., (2002). A Direct Approach to False Discovery Rates. Journal of the Royal Statistical Society. Series B (Statistical Methodology). 64: 479.

Storey, J., (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. The Annals of Statistics. 31: 2013–2035.

Storey, J. and R. Tibshirani (2003a). SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays, in: Parmigiani, G., Garrett, E.S., Irizarry, R.A., Zeger, S.L. (Eds.), The Analysis of Gene Expression Data: Methods and Software, Statistics for Biology and Health. Springer, New York, NY, pp. 272–290.

Storey, J. and R. Tibshirani (2003b). Statistical significance for genomewide studies. Proc Natl Acad Sci U.S.A. 100: 9440–5.

Storey, J.D., A.J. Bass, A. Dabney, D. Robinson and G. Warnes (2020). qvalue: Q-value estimation for false discovery rate control. Bioconductor version: Release (3.10).

Tamhane, A.C. and J. Gou (2018). Advances in p-Value Based Multiple Test Procedures. Journal of Biopharmaceutical Statistics. 28: 10–27.

Tibshirani, R., M.J. Seo, G. Chu, B. Narasimhan and J. Li (2018). Package 'samr': Significance Analysis of Microarrays for differential expression analysis, RNAseq data and related problems. Version 3.0.

Tusher, V.G., R. Tibshirani and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences. 98: 5116–5121.

Vovk, V. and R. Wang (2019a). Combining e-values and p-values (SSRN Scholarly Paper No. ID 3504009). Social Science Research Network, Rochester, NY.

Vovk, V. and R. Wang (2019b). True and false discoveries with e-values. arXiv preprint arXiv:1912.13292.

Westfall, P.H. and S.S. Young (1993). Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, New York.

Westfall, P.H. and R.D. Wolfinger (1997). Multiple Tests with Discrete Distributions. The American Statistician. 51: 3–8.

Westfall, P.H. and J.F. Troendle (2008). Multiple Testing with Minimal Assumptions. Biometrical Journal. 50: 745–755.

Wright, S.P., (1992). Adjusted P-Values for Simultaneous Inference. Biometrics. 48: 1005–1013.

Xia, Y., L. Li, S.N. Lockhart and W.J. Jagust (2019). Simultaneous Covariance Inference for Multimodal Integrative Analysis. Journal of the American Statistical Association. 0: 1–13.

Zaykin, D.V., S.S. Young and P.H. Westfall (2000). Using the false discovery rate approach in the genetic dissection of complex traits: a response to Weller et al. Genetics. 154: 1917–8.

Zhu, Y. and W. Guo (2020). Family-wise error rate controlling procedures for discrete data. Stat. Biopharm. Res. 12: 117–128.