Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Tutorial in biostatistics: multiple hypothesis testing in genomics

Jelle J. Goeman[a]*, Aldo Solari[b]

This paper presents an overview of the current state-of-the-art in multiple testing in genomics data from a user's perspective. We describe methods for familywise error control, false discovery rate control and false discovery proportion estimation and confidence, both conceptually and practically, and explain when to use which type of error rate. We elaborate the assumptions underlying the methods, and discuss pitfalls in the interpretation of results. In our discussion we take into account the exploratory nature of genomics experiments, looking at selection of genes before or after testing, and at the role of validation experiments. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: FDR; false discovery proportion; false positives; Bonferroni

1. Introduction

In modern molecular biology, a single researcher often performs hundreds or thousands of times more hypothesis tests in an afternoon than researchers from a previous generation performed in a lifetime. It is no wonder, therefore, that the methodological discussion in this field has quickly moved from the question whether to correct for multiple testing to the question how to correct for it. In fact, the scale of multiple hypothesis testing problems in genomics experiments is enormous, with numbers of tests ranging from dozens or hundreds in high-throughput screens, to tens or hundreds of thousands in gene expression microarrays or genome-wide association studies, and even to several millions in modern next generation sequencing experiments. The huge scale of these studies, together with the exploratory nature of the research questions, makes the multiple testing problems in this field different from multiple testing problems traditionally encountered in other contexts. Many novel multiple testing methods have been developed in the last two decades, and traditional methods have been reappraised. This paper aims to give an overview of the current state-of-the-art in the field, and to give guidelines to practitioners faced with large exploratory multiple testing problems.

Earlier reviews on multiple testing that deal with genomics have appeared. We mention especially the excellent review by Dudoit, Shaffer and Boldrick [1], and the retrospective overview by Benjamini [2]. More technical overviews can be found in the papers of Farcomeni [3] and Roquain [4] and the book by Dudoit and Van der Laan [5].

1.1. Why multiple testing?

Hypothesis tests are widely used as the gatekeepers of the scientific literature. In many fields, scientific claims are not believed unless corroborated by rejection of some hypothesis. Hypothesis tests are not free of error, however, and for every hypothesis test there is a risk of falsely rejecting a hypothesis that is true, i.e. a type I error, and of failing to reject a hypothesis that is false, i.e. a type II error. Type I errors are traditionally considered more problematic than type II errors. If a rejected hypothesis allows publication of a scientific finding, a type I error

[a] *Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands.*
[b] *Department of Statistics, University of Milano-Bicocca, via Bicocca degli Arcimboldi 8, 20126 Milan, Italy.*
* *Correspondence to: Jelle Goeman, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Medical Statistics (S5-P), Postbus 9600, 2300 RC Leiden, The Netherlands. E-mail: j.j.goeman@lumc.nl*

brings a “false discovery”, and the risk of publication of a potentially misleading scientific result. Type II errors, on the other hand, mean missing out on a scientific result. Although unfortunate for the individual researcher, the latter is, in comparison, less harmful to scientific research as a whole.

In hypothesis tests the probability of making a type I error is bounded by α, an ‘acceptable’ risk of type I errors, conventionally set at 0.05. Problems arise, however, when researchers do not perform a single hypothesis test but many of them. Since each test again has a probability of producing a type I error, performing a large number of hypothesis tests virtually guarantees the presence of type I errors among the findings. As the type I errors among the findings are likely to be the most surprising and novel ones, they have a high risk of finding their way into publications.

The key goal of multiple testing methods is to control, or at least to quantify, the flood of type I errors that arise when many hypothesis tests are performed simultaneously. Different methods do this in different ways, as there are different ways to generalize the concept of type I error to the situation with more than one hypotheses, as we’ll see in Section 1.3.

It is helpful to see the problem of multiple testing as a problem caused by selection [6, 7]. Although even without multiple testing correction the probability of a type I error in each individual hypothesis remains equal to α regardless of the number of hypotheses that have been tested, the researcher will tend to emphasize only the rejected hypotheses. These rejected hypotheses are a selected subset of the original collection of hypotheses, and type I errors tend to be overrepresented in this selection. The probability of a selected hypothesis to be a type I error is therefore much larger than α. Multiple testing methods aim to correct for this selection process and bring type I error probabilities back to α even for selected hypotheses. Different types of multiple testing methods do this in different ways.

The same type of selection problem occurs whenever many hypotheses are tested, and only a selected subset of those is reported or emphasized [8]. It arises when a single researcher simultaneously tests many genomic markers or probes. It arises when these probes are tested for association with multiple phenotypes of interest. It also arises when many different tests or models are tried on the same data set. It also arises when many research groups are working on the same problem, and only the ones that are successful publish, resulting in *publication bias* [9]. In this paper, we concentrate of the first problem only, because it is most characteristic of genomics data.

A recurring problem in multiple testing is to define what the proper collection, or *family* of hypotheses is over which multiple testing correction needs to be done [10]. As a thought experiment, compare a single researcher performing a genomics experiment with 1,000 probes, or 1,000 researchers each performing the same experiment but with a single probe. In both situations the same multiple testing problem occurs, but only the first case would be treated as one. Conventionally, the family over which multiple testing correction is done is all the hypotheses tested in the analysis leading to a single publication. This is arbitrary but practical, and takes into account most of the selection that is done out of sight of other researchers. We can imagine that if many other research groups tried some experiment and failed, reviewers and readers will have heard about this and would be more skeptical about a similar result when it is finally submitted or published. In genomics, the natural family over which multiple testing is done is the set of hypotheses relating to the same research question, ranging over the probes in the experiment. This is the multiple testing problem we will focus on in this tutorial. If multiple research questions are asked per probe this adds another layer to the multiple testing problem. This problem is beyond the scope of this tutorial, although we will touch upon it briefly in Section 5.2.

1.2. Exploration and validation

In the past, much multiple testing method development has focused on clinical trial applications, in which the number of hypotheses to be tested is limited and carefully selected, and in which type I error control has to be very strict, because the clinical trial is often the last scientific stage before a new treatment is allowed on the market.

Genomics experiments are very different. In a gene expression microarray experiment, for example, we typically want to test for differential expression of the each of the probes on the microarray chip. In this experiment the number of hypotheses numbers in tens or hundreds of thousands. These hypotheses have not been purposefully selected for this experiment, but are simply the ones that are available with the technology used. Moreover, the microarray experiment is often not even the final experiment before publication of the scientific paper. Separate validation experiments usually follow for some or all of the probes found differentially expressed. In many ways, the analysis of genomics experiments resembles exploratory more than confirmatory research. The purpose of the experiment is to come up with a list of promising candidates, to be further investigated by the same research group before publication. These promising candidates are often not only chosen on the basis of p-values or other statistical measures, also using biological considerations. That too is a characteristic of exploratory research.

The traditional view has always been that exploratory research does not require formal hypothesis testing, let alone multiple testing correction. In this view, results of exploratory analysis only need to be suggestive, and providing evidence for the results found is the task of subsequent experiments [10]. This view, in which *anything goes* in exploratory research, turns out to be not completely satisfactory in large-scale genomics experiments [11] for two reasons.

In the first place, it is difficult for a researcher to judge which results stand out as suggestive. A plot of the top ranking result out of tens of thousands will always look impressive, even when the data are pure noise. Before venturing into validation experiments that involve an investment of time and money, researchers like to be assured that they are not wasting too many resources on chasing red herrings.

Secondly, validation experiments are not always sufficiently independent to bear the full burden of proof for the final findings. We distinguish three types of validation experiments. First, *full replication* is repetition of the findings of the experiment by a different research group using different techniques and new subjects. Second, *biological validation* is a replication of the findings by the same research group, using the same technique or a different one, but using new subjects. Third, *technical validation*, is replication of the findings by the same research group on the same subjects, but using a different technique, e.g. redoing microarray expression measurements by a PCR. Full replication is the best type of validation, but by definition not feasible within a single research group. Biological validation is a good second, and is sufficient as validation, even though some biases inherent in the experimental design may be replicated in the validation experiment, especially if the same techniques are used in exploratory experiment and validation experiment. Technical replication, however, is hardly validation at all. Any type I errors coming up in the exploratory experiment are likely to be replicated exactly in a technical validation, as the same subjects will typically show the same patterns if measured twice. If, as often happens for practical reasons, a technical validation is all that is available, then the burden of proof for the final results rests in the full genomics experiment, and rigorous multiple testing correction in that experiment is the only way to prevent false positive findings.

If more than one finding is to be validated in a validation experiment, the results of the validation experiment, of course, require multiple testing correction to prevent type I errors. Validation experiments are not exploratory but confirmatory experiments, and should be treated as such.

1.3. Concepts and outline

There are many ways of dealing with type I errors. In this tutorial we focus on three types of multiple testing methods: those that control the familywise error (FWER), those that control the false discovery rate (FDR), and those that estimate the false discovery proportion (FDP) or make confidence intervals for it. We start by clarifying the terms involved. Methods for FWER control are discussed in Section 2, and methods for FDR control in Section 3. FDP estimation and confidence is treated in Section 4.

The multiple testing problems we consider in this paper have a simple structure. We have a collection $\mathcal{H} = (H_1, \ldots, H_m)$ of hypotheses of interest. An unknown subset $\mathcal{T} \subseteq \mathcal{H}$ of size m_0 of these hypotheses is true, while the remaining collection $\mathcal{F} = \mathcal{H} \setminus \mathcal{T}$ of size $m_1 = m - m_0$ are false. On the basis if the data our goal is to choose a subset $\mathcal{R} \subseteq \mathcal{H}$ of hypotheses to reject. We try to let this set $\mathcal{R}$ coincide with the set $\mathcal{F}$ as much as possible. Two types of error can be made: false positives, or type I errors, are the rejected hypotheses that are not false, i.e. $\mathcal{R} \cap \mathcal{T}$; false negatives or type II errors are the false hypotheses that we failed to reject, i.e. $\mathcal{F} \setminus \mathcal{R}$. Rejected hypotheses are sometimes called *discoveries*, and the terms *true discovery* and *false discovery* are sometimes used for correct and incorrect rejections.

We can summarize the numbers of errors occurring in a hypothesis testing procedure in a contingency table such as Table 1. We can observe m and $R = \#\mathcal{R}$, but all quantities in the first two columns of the table are unobservable.

	true	false	total
rejected	V	U	R
not rejected	$m_0 - V$	$m_1 - U$	$m - R$
total	m_0	m_1	m

Table 1. Contingency table for multiple hypothesis testing: rejection versus truth or falsehood of hypotheses.

Type I and type II errors are in direct competition with each other, and choosing a set $\mathcal{R}$ that has fewer type I errors usually results in more type II errors. Focus in multiple testing is on keeping small either on the number V of type I errors or the *false discovery proportion* Q, defined as

$$Q = \frac{V}{\max(R, 1)} = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{otherwise,} \end{cases}$$

which is the proportion of false rejections among all rejections made.

FWER control and FDR control methods choose the set $\mathcal{R}$ of rejected hypotheses as a function of the data, typically of the form $\mathcal{R} = \{H_i : p_i \leq T\}$ for some data-dependent threshold T, where $p_1, \ldots, p_m$ are p-values corresponding to hypotheses $H_1, \ldots, H_m$. Because the p-values and the threshold T are random, the set $\mathcal{R}$ is random and so are V and Q. If both V and Q are random variables, we cannot keep the values of V and Q themselves small, but must focus on relevant aspects of the distributions of these variables. FWER and FDR controlling methods focus on different summaries of the distribution of V and Q, namely FWER, given by

$$\mathrm{FWER} = \mathrm{P}(V > 0) = \mathrm{P}(Q > 0)$$

and FDR [12], given by

$$\mathrm{FDR} = \mathrm{E}(Q).$$

FWER focuses on the probability that the rejected set contains any error, while FDR looks at the expected proportion of errors among the rejections. Controlling FWER or FDR at level α means that the set $\mathcal{R}$ (i.e. the threshold T) is chosen in such a way that the corresponding aspect of the distribution of Q is guaranteed to be at most α. FWER and FDR are by no means the only aspects of the distribution of Q or V that can be of interest, only the most popular ones. Several authors have proposed other summaries, too many to mention all of them here [2].

The two error rates FDR and FWER are related. Since $0 \leq Q \leq 1$, we have $\mathrm{E}(Q) \leq \mathrm{P}(Q > 0)$, which implies that every FWER-controlling methods is automatically also an FDR-controlling methods. Since FDR is smaller than FWER, it is easier to keep the FDR below a level α than to keep the FWER below the same level, and we can generally expect FDR-based method to have more power than FWER-based ones. Under the complete null hypothesis (i.e. if $m_1 = 0$), Q is a Bernoulli variable and FDR and FWER are identical.

To understand FDR it is helpful to look at the contingency Table 1 from the analogy of a contingency table in clinical testing. If we equate a rejected hypothesis with a positive result from a clinical test, then the expected ratio $\mathrm{E}(V/m_0)$ corresponds to 1 minus the specificity of the test, and, if always $R > 0$, the FDR $\mathrm{E}(Q)$ corresponds to 1 minus the positive predictive value [13]. Unadjusted hypothesis testing guarantees that $\mathrm{E}(V/m_0)$ stays below α, and therefore keeps the specificity in Table 1 above $1 - \alpha$. It is known, however, that at low prevalence m_1/m, high specificity can still coincide with a low positive predictive value. For this reason, methods that control FDR tend to be much stricter than unadjusted testing if the prevalence m_1/m is low. It is important to realize, however, that the converse is also true: at high prevalence, high positive predictive value can coincide with low specificity. Consequently, control of FDR does not necessarily imply type I error control for individual tests. For example, in an extreme situation, if it is known a priori that $m_0/m \leq \alpha$, we may reject all hypotheses and still control FDR. It is clear that this procedure does not keep type I error for each individual hypothesis. It is therefore a reasonable additional demand of methods that control FDR that they also keep *per comparison* type I error for individual hypotheses, and not reject hypotheses with p-values above α, but many of them do this. FWER, by keeping the numerator V of both FDR and individual type I error small, keeps both specificity and positive predictive value maximal.

Related to the contingency table view of Table 1 is the empirical Bayes view of FDR. In this view, the truth or falsehood of each hypothesis is not seen as fixed, but as random, with the indicator of each hypothesis' truth a Bernoulli variable with common success probability π_0. Under this additional assumption all quantities in Table 1 become random, and we can legitimately speak about the probability that a hypothesis is true. In this model, the conditional probability that a hypothesis is true given that is has been rejected is closely related to FDR, and is known as the empirical Bayes FDR [13]. We come back to this view of FDR in Sections 4.1 and 4.2.

Both FDR and FWER are proper generalizations of the concept of type I error to multiple hypotheses; if there is only one hypothesis ($m = 1$) the two error rates are identical, and equal to the regular type I error. FDR and FWER generalize type I error in a different way, however. We can say that if FWER of a set of hypotheses $\mathcal{R}$ is below α, then *for every* hypothesis in $H \in \mathcal{R}$ the probability that H is a type I error is below α. FDR control, on the other hand, only implies type I error control *on average* over all hypotheses $H \in \mathcal{R}$. Properties of "for every"-type statements are different from those of "on average"-type statements. In particular, FWER has the *subsetting property* that if a set $\mathcal{R}$ of hypotheses is rejected by an FWER-controlling procedure, then FWER control is also guaranteed for any subset $\mathcal{S} \subset \mathcal{R}$. The corresponding property does not hold for FDR control. In fact, it was argued by Finner and Roters [14] that a procedure that guarantees FDR control not only for the rejected set itself, but also for all subsets, must be an FWER-controlling procedure. While FWER control is a statement that immediately translates to type I error control of individual hypotheses, FDR control is only a statement on the full set $\mathcal{R}$, and one which does not translate to subsets of $\mathcal{R}$ or individual hypotheses in $\mathcal{R}$. This subsetting property, or lack of it, has implications for the way FWER and FDR can be used, and we come back to this in Sections 2.5 and 3.4.

Methods that control an error rate, such as FDR or FWER, contrast with methods that estimate the number or proportion of errors. Such estimation methods are not interested in the distribution of V or Q induced by a distribution of the rejected set $\mathcal{R}$, but only in the actual value of V or Q realized by a specific rejected set $\mathcal{R}$. For a specific non-random set $\mathcal{R}$, the value of the FDP Q is a fixed but unknown quantity, a function of the sets $\mathcal{T}$ and $\mathcal{R}$. This quantity may be estimated, and confidence intervals can be constructed for it. Making such estimates and confidence statements for the value of Q in the eventually chosen rejected set is the goal of FDP estimation methods. In practice, of course, the set $\mathcal{R}$ to be rejected is not determined before data collection, but will be chosen in some data-dependent way. Any estimates and confidence statements for Q need to be corrected for bias resulting from such a data-dependent choice. We will look into FDP estimation methods in greater detail in Section 4.

1.4. Accounting for dependence of p-values

In statistics, stronger assumptions generally allow more powerful statements. In multiple testing, the most crucial assumptions to be made concern the dependence of the p-values of the different hypotheses. Much work has been done under the assumption of independence of p-values, but this work is of little practical value in genomics data, in which molecular measurements typically exhibit strong but a priori unknown correlations. Methods with more realistic assumptions come in three major flavors. The first kind makes no assumptions at all. They protect against a 'worst case' dependence structure, and are conservative for all other dependence structures. The second kind gains power by assuming that the dependence structure of the p-values is such that Simes' inequality holds. The third kind uses permutations to adapt to the dependence structure of the p-values. Since all three types of assumptions are used in methods for FWER control, FDR control and FDP estimation, we discuss the underlying assumptions in detail before we move on to specific methods.

All methods we consider in this tutorial start from a collection of test statistics $S_1, \ldots, S_m$, one for each hypothesis tested, with corresponding p-values $p_1, \ldots, p_m$. We call these p-values *raw* as they have not been corrected for multiple testing yet. By the definition of a p-value, if their corresponding null hypothesis is true, these p-values are either uniformly distributed between 0 and 1 or they are stochastically smaller than that, i.e. we have

$$\mathrm{P}(p_i \leq t) \leq t. \tag{1}$$

In practice, raw p-values are often only approximate, as they are derived through asymptotic arguments or other approximations. It should always be kept in mind that such asymptotic p-values can be quite inaccurate, especially for small sample sizes, and that their relative accuracy decreases when p-values become smaller.

Methods that make no assumptions on the dependence structure of p-values are based on some probability inequality. Two such inequalities are relevant for methods described in this tutorial. The first is the Bonferroni inequality, which is discussed in detail in Section 2.1. The second is an inequality due to Hommel, which states that with probability at least $1 - \alpha$, we have that simultaneously

$$q_{(i)} > \frac{i\alpha}{m_0 \sum_{j=1}^{m_0} 1/j} \qquad \text{for all } i = 1, \ldots, m_0, \tag{2}$$

where $q_{(1)} \leq \ldots \leq q_{(m_0)}$ are the m_0 ordered p-values of hypotheses corresponding to true null hypotheses. Hommel's inequality is valid whatever the dependence of p-values, as long as (1) holds. The difference between the series $\sum_{j=1}^{m_0} 1/j$ appearing in the denominator and $\log(m_0)$ converges to the Euler-Mascheroni constant $\gamma \approx 0.577$ as $m_0 \to \infty$.

Probability inequalities have a 'worst case' distribution for which the inequality is an equality, but are are strict inequalities for most distributions. Multiple testing methods based on such inequalities are therefore conservative for all p-value distributions except for this 'worst case'. Such 'worst case' distributions are often quite unrealistic, and this is especially true for Hommel's inequality [15], which can be quite conservative for more realistic distributions. The worst case of the Bonferroni inequality is discussed in Section 2.1. Assumption-free methods discussed in this tutorial are the Bonferroni and Holm methods (Sections 2.1 and 2.2) for FWER control, using the Bonferroni inequality, the Benjamini & Yekutieli method (Section 3.2) for FDR control, related to Hommel's inequality, and one of the confidence bound methods for FDP from Section 4.3, also using Hommel's inequality. It is worth mentioning that Hommel's FWER control method (Section 2.3) is unrelated to Hommel's inequality.

To avoid having to cater for exotic worst case distributions, assumptions can be made to exclude them. In particular, a set of assumptions can be made that allows both the Benjamini & Hochberg method for FDR control (Section 3.1) and the related Simes inequality [16]. The latter is a probability inequality related to Hommel's inequality, which says that with probability at least $1 - \alpha$, simultaneously

$$q_{(i)} > \frac{i\alpha}{m_0} \qquad \text{for all } i = 1, \ldots, m_0. \tag{3}$$

The conditions under which the Simes inequality and the Benjamini & Hochberg procedure hold have been extensively studied by Sarkar [17, 18] and others [19]. Both have been proved to hold under the sufficient condition that *positive dependence through stochastic ordering* (PDSS) holds on the test statistics of the subset of true null hypotheses. The same condition is also known under the name of *positive regression dependence on a subset*. Examples of cases under which this condition is true include one-sided test statistics that are marginally normally or t-distributed, if all correlations between test statistics are positive; and two-sided test statistics that are normally or t-distributed under more general assumptions on the correlation matrix. Even though this condition is not guaranteed to hold for all distributions relevant for genomics, Simes' inequality turns out to be quite robust in practice. This has been corroborated theoretically by Rødland [20], who wrote that "distributions for which Simes' procedure fails more dramatically must be somewhat bizarre". It has also been corroborated by many simulation experiments [16, 21, 22, 23] which indicate that the Benjamini & Hochberg procedure and the Simes inequality are highly robust. The general consensus seems to be that in genomics data, especially for the ubiquitous case of two-sided tests that are asymptotically normal, it is safe to assume that Simes inequality and the Benjamini & Hochberg procedure are valid [24].

Simes' inequality is also a strict inequality for some distributions, but the 'worst case', for which the Simes inequality is not conservative, is the case of independent uniform p-values [16], which is relatively unexotic. For the Benjamini & Hochberg procedure, the 'worst case' for which the procedure is least conservative is the *dirac-uniform* configuration in which all false hypotheses always have p-values exactly zero, and p-values of true null hypotheses are independent and uniform [25].

Methods in this tutorial that are based on the assumptions leading to Simes' inequality are Hochberg's and Hommel's methods for FWER control (Section 2.3), the procedure of Benjamini & Hochberg for FDR control (Section 3.1), and the Simes-based confidence method for FDP (Section 4.3).

Another way to deal with the unknown dependence structure of p-values is permutation testing. This is a large subject by itself, which we cannot hope to cover fully in this tutorial: we focus only on application of permutations for multiple testing. Readers unfamiliar with the basics or fundamentals of permutation testing are referred to the books by Good [26] and Pesarin [27]. Permutation tests have two great advantages, both of which translate to permutation-based multiple testing. First, they give exact error control without relying on the assumption (1), allowing reliable testing even when asymptotic p-values are unreliable. Second, permutation tests do not use any probability inequality, but attain the exact level α regardless of the distribution of the underlying data. Permutation-based multiple testing is said to 'adapt' to the dependence structure of the raw p-values. It is not conservative for any dependence structure of the p-values, and it can be especially powerful in case of strong dependence between p-values.

Unfortunately, valid permutation tests are not defined for all null hypotheses in all models; permutation tests are not assumption-free [28]. A valid permutation is a *null-invariant* transformation of the data, which means that it does not change the joint distribution of the p-values corresponding to true hypotheses [29]. For example, in a genomic case-control study, if we can assume that the joint distribution of the genomics measurements corresponding to true hypotheses is identical between cases and controls, the classical data transformation that randomly reassigns case and control labels to the subjects is a valid permutation. The same permutation is not a valid permutation if we are not willing to make that assumption of identical joint distributions in cases and controls. For example, if measurements of cases and controls corresponding to true hypotheses can have different variance or if the correlation structure of these measurements may differ between cases and controls, the same permutations do not lead to valid permutation tests. In fact, in such a model no valid permutation test exists, and asymptotic methods must be used. Generally, only relatively simple experimental designs allow permutation tests to be used.

Multiple testing methods based on permutations include the methods of Westfall & Young for FWER control (Section 2.4) and Meinshausen's method for FDP confidence (Section 4.4). An exact and powerful permutation-based method for FDR control has not yet been found, but we will discuss developments in this area in Section 3.3.

In rare cases, important aspects of the correlation structure of the test statistics or p-values can be assumed to be known a priori. Such information can then be exploited to obtain more powerful multiple testing adjustment. In genome-wide association studies, for example, it has been asserted that the correlation structure between the p-values in such a study is such that a fixed genome-wide significance level of $\alpha = 5 \times 10^{-8}$, corresponding to a Bonferroni adjustment for 10^6 tests, is sufficient for FWER control, however many hypotheses are tested [30]. The validity of such a universal threshold of course depends crucially on the validity of the assumption on the correlation structure of the underlying p-values [31].

1.5. Recurrent examples

To illustrate all the different methods in this tutorial we use two gene expression microarray data sets, both with a survival phenotype. The first one is the data of Van de Vijver [32]. This data set has gene expression profiles of 4,919 probes for 295 breast cancer patients. The second data set, of Rosenwald [33], has gene expression profiles of 7,399 probes for 240 diffuse B-cell lymphoma patients. Median follow up was 9 years for the breast cancer patients, and 8 years for the lymphoma patients. Although both data sets are by now a decade old, and new technologies, such as RNA sequencing, have since appeared, the multiple testing issues at stake have not changed; only the scale of problems has increased further. These data sets still serve very well for illustration purposes.

In each data set we performed a likelihood ratio test in a Cox proportional hazards model for each probe, testing association of expression with survival. The plot of the sorted p-values and a histogram of the p-values are given in Figure 1. From the left-hand plot we may suspect that many of the hypotheses that are tested in the Van de Vijver data are false. If the overwhelming majority of the hypotheses were true, we would expect the histogram of the p-values to be approximately uniform and the plot of the sorted p-values approximately to follow the dotted line, because p-values of true null hypotheses follow a uniform distribution. There is less immediate evidence for differential expression in the Rosenwald data, although there seems to be enrichment of low p-values here too. Before jumping to conclusions, however, it should be noted that, although we expect uniform profiles for the p-values of true null hypotheses on average, correlations between p-values can make figures such as the ones in Figure 1 highly variable. Without further analysis, it is not possible to attribute the deviations from uniformity in these plots confidently to the presence of false hypotheses.

1.6. Available software

Since R is the dominant statistical software package for the analysis of genomics data, we concentrate on available packages for that program. Some of the less complex methods can easily be perfomed ‘by hand’ in a spreadsheet, and we summarize these methods in the box of Algorithm 1. A single useful and user-friendly suite of methods for multiple testing that encompasses many more methods than we can discuss here is available though the R package of the μTOSS project [34]. For those who do not use R, SAS also has a large collection of multiple testing procedures. Multiple testing in other commercial packages such as SPSS and Stata is, unfortunately, very limited.

2. Methods for FWER control

The traditional and most well-known method to control FWER is the method of Bonferroni, which replaces the cut-off α for declaring significance of individual tests by α/m. Despite being so well-known, or perhaps because of this, there is a lot of misunderstanding about the method of Bonferroni in the literature. We will start this section with a discussion of these misunderstandings, before we move on to the more powerful methods of Holm, Hochberg, Hommel and Westfall & Young.

2.1. Bonferroni

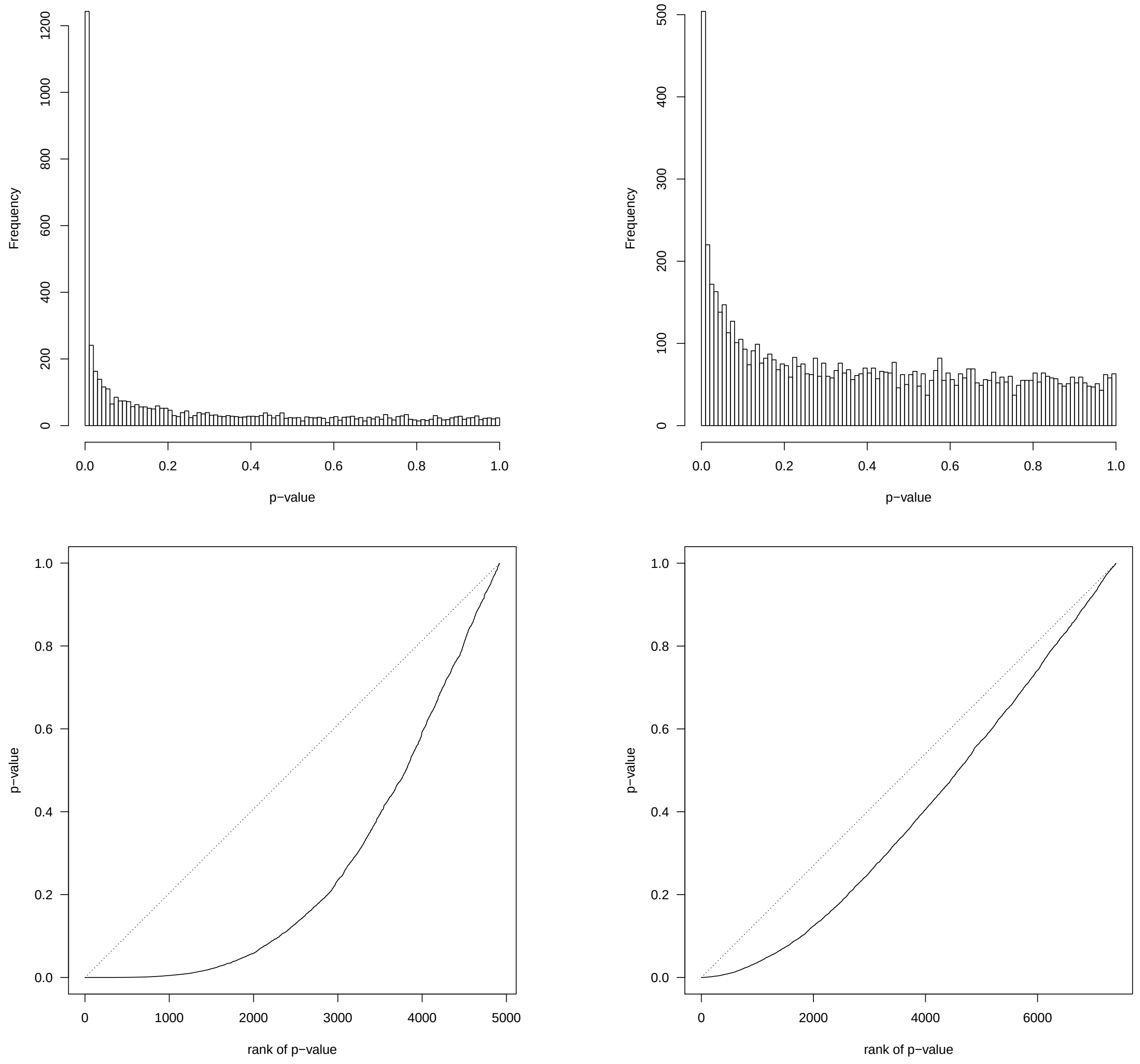
One of the most widespread misunderstandings of the method of Bonferroni is that it would be based on an assumption of independence between p-values [35]. Indeed, the probability of making a false rejection if all m_0 p-values of true null hypotheses are independent, and we perform each test at level α/m is $1-(1-\alpha/m)^{m_0}$. Expanding this expression, the first term, which is dominant for small α or large m, is $m_0\alpha/m \leq \alpha$. This reasoning, often presented as motivation for the Bonferroni procedure, does not do it justice. It makes Bonferroni seem like a method that only provides approximate FWER control, and that requires an assumption of independence for its validity.

In fact, the Bonferroni method is a corollary to Boole’s inequality, which says that for any collection of events $E_1, \ldots, E_k$, we have

$$\mathrm{P}\big(\bigcup_{i=1}^{k} E_i\big) \leq \sum_{i=1}^{k} \mathrm{P}(E_i).$$

It follows from Boole’s inequality together with (1) that, if $q_1, \ldots, q_{m_0}$ are the p-values of the true null hypotheses, that the probability that there is some i for which $q_i \leq \alpha/m$ is given by

$$\mathrm{P}\big(\min_i q_i \leq \alpha/m\big) = \mathrm{P}\big(\bigcup_{i=1}^{m_0}\{q_i \leq \alpha/m\}\big) \leq \sum_{i=1}^{m_0} \mathrm{P}(q_i \leq \alpha/m) \leq m_0\frac{\alpha}{m} \leq \alpha. \tag{4}$$



(a) Van de Vijver data

(b) Rosenwald data

Figure 1. Histogram and profile of p-values from the data sets of Van de Vijver and Rosenwald.

A few things can be learnt from this derivation. In the first place, the FWER control of Bonferroni is exact, not approximate, and it is valid for all dependence structures of the underlying p-values. Secondly, the three inequalities in (4) indicate in which cases the Bonferroni method can be conservative. The right-hand one shows that Bonferroni does not control the FWER at level α but actually at the stricter level $\pi_0\alpha$, where $\pi_0 = m_0/m$. If there are many false null hypotheses, Bonferroni will be conservative. The middle inequality, that uses (1), says thay Bonferroni is conservative if the raw p-values are. The left-hand inequality is due to Boole's law. This inequality is a strict one in all situations except the one in which all events $\{q_i \leq \alpha/m\}$ are disjoint. From this, we conclude that Bonferroni is conservative in all situations except in the situation that the rejection events of the true hypotheses are perfectly negatively associated.

The conservativeness of Bonferroni in situations in which Boole's inequality is strict deserves more detailed attention. With independent p-values, this conservativeness is present but very minor. To see this we can compare the Bonferroni critical value α/m with the corresponding Sidak [36] critical value $1-(1-\alpha)^{1/m}$ for independent p-values. For $m = 5$ and $\alpha = 0.05$ we find a critical value of 0.01021 for Sidak against 0.01 for Bonferroni. As m increases, the ratio between the two increases to a limit of $-\log(1-\alpha)/\alpha$, which evaluates to only 1.026 for

$\alpha = 0.05$. Much more serious conservativeness can occur if p-values are positively correlated. For example, in the extreme case that all p-values are perfectly positively correlated, FWER control could already have been achieved with the unadjusted level α, rather than α/m. Less extreme positive associations between p-values would also allow less stringent critical values, and Bonferroni can be quite conservative in such situations.

A second, less frequent misunderstanding about Bonferroni is that it would only protect in the situation of the global null hypotheses, i.e. the situation that $m_0 = m$ [35, 37]. This type of control is known as *weak* FWER control. On the contrary, as we can see from (4) Bonferroni controls the FWER for any combination of true and false hypotheses. This is known a *strong* FWER control. In practice, only strong FWER controlling methods are of interest, and methods with only weak control should, in general, be avoided. To appreciate the difference between weak and strong control, consider a method that, if there is at least one p-value below α/m, rejects all m hypotheses, regardless of the p-values they have. This nonsensical method has weak, but not strong FWER control. Related to weak control methods but less overconfident are global testing methods [38, 39] that test the global null hypothesis that $m_0 = m$. If such a test is significant, one can confidently make the limited statement that at least one false hypotheses is present, but not point to which one. In contrast, methods with strong FWER control also allow pinpointing of the precise hypotheses that are false.

When testing a single hypothesis, we often do not only report whether a hypothesis was rejected, but also the corresponding p-value. By definition, the p-value is the smallest chosen α-level of the test at which the hypothesis would have been rejected. The direct analogue of this in the context of multiple testing is the adjusted p-value, defined as the smallest α level at which the multiple testing procedure would reject the hypothesis. For the Bonferroni procedure, this adjusted p-value is given by $\min(mp_i, 1)$, where p_i is the raw p-value.

2.2. *Holm*

Holm's method [40] is a sequential variant of the Bonferroni method that always rejects at least as much as Bonferroni's method, and often a bit more, but still has valid FWER control under the same assumptions. From this perspective, there is no reason, aside from possibly simplicity, to use Bonferroni's method in preference to Holm's.

Holm remedies part of the conservativeness in the Bonferroni method arising from the right-hand inequality of (4), which makes Bonferroni control FWER at level $\pi_0\alpha$. It does that by iterating the Bonferroni method in the following way. In the first step, all hypotheses with p-values at most α/h_0 are rejected, with $h_0 = m$ just like in the Bonferroni method. Suppose this leaves h_1 hypotheses unrejected. Then, in the next step, all hypotheses with p-values at most α/h_1 are rejected, which leaves h_2 hypotheses unrejected, which are subsequently tested at level α/h_2. This process is repeated until either all hypotheses are rejected, or until a step fails to result in any additional rejections. Holm gave a very short and elegant proof that this procedure controls the FWER in the strong sense at level α. This proof is based on Boole's inequality just like that of the Bonferroni method, and consequently makes no assumptions whatsoever on the dependence structure of the p-values.

It is immediately obvious that Holm's method rejects at least as much as Bonferroni's and possibly more. The gain in power is greatest in the situation that many of the tested hypotheses are false, and when power for rejecting these hypotheses is good. Rejection of some of these false hypotheses in the first few steps of the procedure may lead to an appreciable increase in the critical values for the remaining hypotheses. Still, unless the proportion of false hypotheses in a testing problem is very large, the actual gain is often quite small. We can see this in the example data sets of Rosenwald and Van de Vijver. In de Van de Vijver data, the Bonferroni method rejects 203 hypotheses at a critical value of $0.05/4919 = 1.02 \times 10^{-5}$. This allows the critical value in the second step of Holm's procedure to be adjusted to $0.05/4716 = 1.06 \times 10^{-5}$, which allows 3 more rejections. The increase in the critical value resulting from these three rejections is not sufficient to allow any additional rejections, giving a total of 206 rejections for Holm's procedure. In the Rosenwald data, Bonferroni allows only 4 rejections at its critical value of $0.05/7399 = 6.76 \times 10^{-6}$, but the resulting increase in the critical value in Holm's method to $0.05/7395 = 6.76 \times 10^{-6}$ is insufficient to make a difference.

An alternative way of describing Holm's method is via the ordered p-values $p_{(1)}, \ldots, p_{(m)}$. Comparing each p-value $p_{(i)}$ to its corresponding critical value $\alpha/(m - i + 1)$, Holm's method finds the smallest j such that $p_{(j)}$ exceeds $\alpha/(m - j + 1)$, and subsequently rejects all $j - 1$ hypotheses with a p-value at most $\alpha/(m - j)$. If no such j can be found, all hypotheses are rejected.

Adjusted p-values for Holm's method can be calculated using the simple algorithm of Table 1. Because increasing the level of α, when this causes one rejection, may immediately trigger a second one because of the resulting increase in the critical value, it is possible for adjusted p-values in Holm's method to be equal to each other even when the raw p-values are not. The same feature occurs in almost all of the other methods described below.

```
Start with $p$-values for $m$ hypotheses
  1. Sort the $p$-values $p_{(1)}, \dots, p_{(m)}$.
  2. Multiply each $p_{(i)}$ by its adjustment factor $a_i$, $i = 1, \dots, m$, given by
      (a) Holm or Hochberg: $a_i = m - i + 1$
      (b) Benjamini & Hochberg: $a_i = m/i$
      (c) Benjamini & Yekutieli: $a_i = lm/i$, with $l = \sum_{k=1}^{m} 1/k$
  3. If the multiplication in step 2 violate the original ordering, repair this.
      (a) Step-down (Holm): Increase the smallest $p$-value in all violating pairs:
              $\tilde{p}_{(i)} = \max_{j=1,\dots,i} a_j p_{(j)}$
      (b) Step-up (all others): Decrease the highest $p$-value in all violating pairs:
              $\tilde{p}_{(i)} = \min_{j=i,\dots,m} a_j p_{(j)}$
  4. Set $\tilde{p}_{(i)} = \min(\tilde{p}_{(i)}, 1)$ for all $i$.
```

Algorithm 1: Calculating adjusted p-values for the methods of Holm, Hochberg, Benjamini & Hochberg (BH), and Benjamini & Yekutieli (BY). The algorithms above are easy to implement in any spreadsheet program. In R, it is easier to just use the `p.adjust` function, which also has Hommel's method.

2.3. Hochberg and Hommel

Bonferroni's and Holm's methods make no assumptions on the dependence structure of the p-values, and protect against the 'worst case' according to Boole's inequality, which is that the rejection regions of the different tests are disjoint. If we are willing to make assumptions on the joint distribution of the p-values, it becomes possible exclude this worst case a priori, and as a result gain some power.

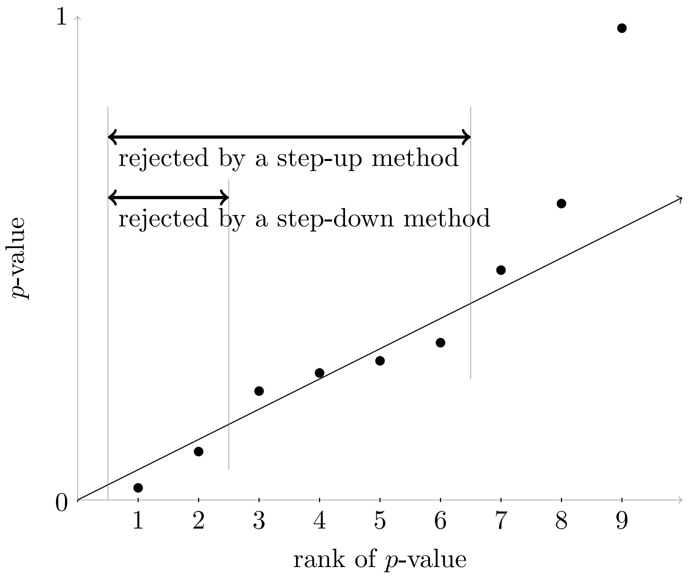
One such assumption could be that the PDSS condition holds for the subset of true hypotheses (Section 1.4). This assumption makes the use of Simes inequality (3) and therefore the use of Hochberg's method [41] possible, which is very similar to Holm's method but more powerful. Hochberg's method (not to be confused with Benjamini & Hochberg's method, Section 3) compares each ordered p-value $p_{(i)}$ to a critical value $\alpha/(m-i+1)$, the same as Holm's. It then finds the largest j such that $p_{(j)}$ is smaller than $\alpha/(m-j+1)$, and subsequently rejects all j hypotheses with p-values at most $\alpha/(m-j+1)$. In the jargon of multiple testing methods, Holm's method is known as a *step-down* method and Hochberg's as its *step-up* equivalent. An illustration of the difference between step-up and step-down methods is given in Figure 2. Holm's method uses the first crossing point between the p-value curve $p_{(1)}, \dots, p_{(m)}$ and the critical value curve $\alpha/m, \alpha/(m-1), \dots, \alpha$ as its critical value, while Hochberg's uses the largest crossing point instead. The step-up/step-down parlance can be somewhat confusing, as Holm's method "steps up" from the smallest p-value to find the crossing while Hochberg's "steps down" from the largest one, but the terminology was originally formulated in terms of test statistics rather than p-values. Comparing to Holm's method, it is clear that Hochberg's method rejects at least as much as Holm's method, and possibly more. If the p-value and critical value curves never cross or only once, Holm's and Hochberg's methods reject the same number of hypotheses. If the same curves cross multiple times or if the smallest p-value is larger than its corresponding critical value, Hochberg's method rejects more.

Hochberg's method is a special case of the powerful closed testing procedure [42], and the proof of its FWER control is based on combining that procedure with the Simes inequality. Hommel [43] showed that a more powerful, although more complicated and computationally more demanding procedure can be constructed from the same ingredients. Hommel's resulting procedure rejects at least as much as Hochberg's, but possibly more [44]. On modern computers the additional computational burden of Hommel's procedure is not an impediment anymore, and Hommel's procedure should always be preferred to Hochberg's just like Holm's procedure should always be preferred to Bonferroni's.

A gain in power of Hochberg's and Hommel's methods over Holm's method can be expected in the situation that a large proportion of the hypotheses is false, but the corresponding tests have relatively low power, or if there are positive associations between the p-values of these false hypotheses. In practice, just like the gain from Bonferroni to Holm's method, the number of additional rejections allowed by Hochberg or Hommel is often small. In the Van de Vijver data set, the curve of the ordered p-values crosses the curve of the Holm and Hochberg critical values only once, so the number of rejections is identical to 206 in both methods. Hommel's method, however, is able to improve upon this by a further three rejections, making a total of 209. In the Rosenwald data, neither method is able to improve upon the 4 rejections that were already found by the Bonferroni procedure.

The algorithm for calculating adjusted p-values in Hochberg's method is given in Table 1. For Hommel's method this calculation is less straightforward, and we refer to the `p.adjust` function in `R`. Step-up methods tend to give more tied adjusted p-values than step-down ones, and may sometimes return long lists of identical adjusted p-values.

Figure 2. Comparison of rejections by step-up and step-down methods with the same critical values. The dots are observed ranked p-values. The line represents the critical values. Step-down methods reject all hypotheses up to, but not including, the first p-value that is larger than its critical value. Step-up methods reject all hypotheses up to and including the last p-value that is smaller than its critical value.



2.4. Permutations and Westfall & Young

Instead of making assumptions on the dependence structure of the p-values, it is also possible to adapt the procedure to the dependence that is observed in the data by replacing the unknown true null distribution with a permutation null distribution. In FWER control, the most relevant aspect of the unknown null distribution is the α-quantile of the distribution of the minimum p-value of the m_0 true hypotheses. This is the quantile that Bonferroni bounds from below by α/m. The method of Westfall & Young [45] uses permutations to obtain a more accurate threshold. It isshown to be asymptotically optimal for a broad class of correlation structures [46]. Two variant's of Westfall & Young's methods exist: the max$T$ and min$P$ methods.

In the Van de Vijver or Rosenwald data, a permutation can be a reallocation of the survival time and status to the subjects, so that each subject's gene expression vector now belongs to the survival time of a different subject. For the probes for which there is no association between survival time and gene expression, we can assume the distribution of this permuted data set to be identical to that of the original, thus satisfying the null invariance condition. Since permutation-testing is not assumption-free, it is important to check carefully that permutations are indeed valid. If an additional covariate would be present in the Cox proportional hazards model, for example, the survival curves would not have been identical between individuals, and null invariance would have been much more problematic. In situations in which null invariance is not satisfied, simple per-hypothesis permutation testing combined with Holm's or Hommel's method can sometimes be an alternative to Westfall & Young.

Practically, the max$T$ method of Westfall & Young, applied to the p-values, starts by making k permuted data sets, and recalculating all m p-values for each permuted data set. Let's say we store the results in an $m \times k$ matrix $\mathbf{P}$. We find the k minimal p-values along each column to obtain the permutation distribution of the minimum p-value out of m. The α-quantile $\tilde{\alpha}_0$ of this distribution is the permutation-based critical value, and Westfall & Young reject all hypotheses for which the p-value in the original data set is strictly smaller than $\tilde{\alpha}_0$. Next, we may continue from this result in the same step-down way in which Holm's method continues on Bonferroni's. In the next step we may remove from the matrix **P** all rows corresponding to the hypotheses rejected in the first step, and recalculate the k minimal p-values and their α-quantile $\tilde{\alpha}_1$. Removal of some hypotheses may have increased the values of some of the minimal p-values, so that possibly $\tilde{\alpha}_1 > \tilde{\alpha}_0$. We may now reject any additional hypotheses that have p-values below the new quantile $\tilde{\alpha}_1$. The process of removing rows of rejected hypotheses from **P** and recalculating the α-quantile of the minimal p-values is repeated until any step fails to result in additional rejections, or until all hypotheses have been rejected, just like in Holm's procedure.

Westfall & Young's min$P$ method is similar to the max$T$ method, except that instead of the raw p-values it uses the per-hypothesis permutation p-values in the matrix **P**. A fast algorithm for this procedure was designed by Ge [47]. Since permutation p-values take only a limited number of values, the matrix **P** will always contain many tied values, which is an important practical impediment for the min$P$ method, as we'll see below.

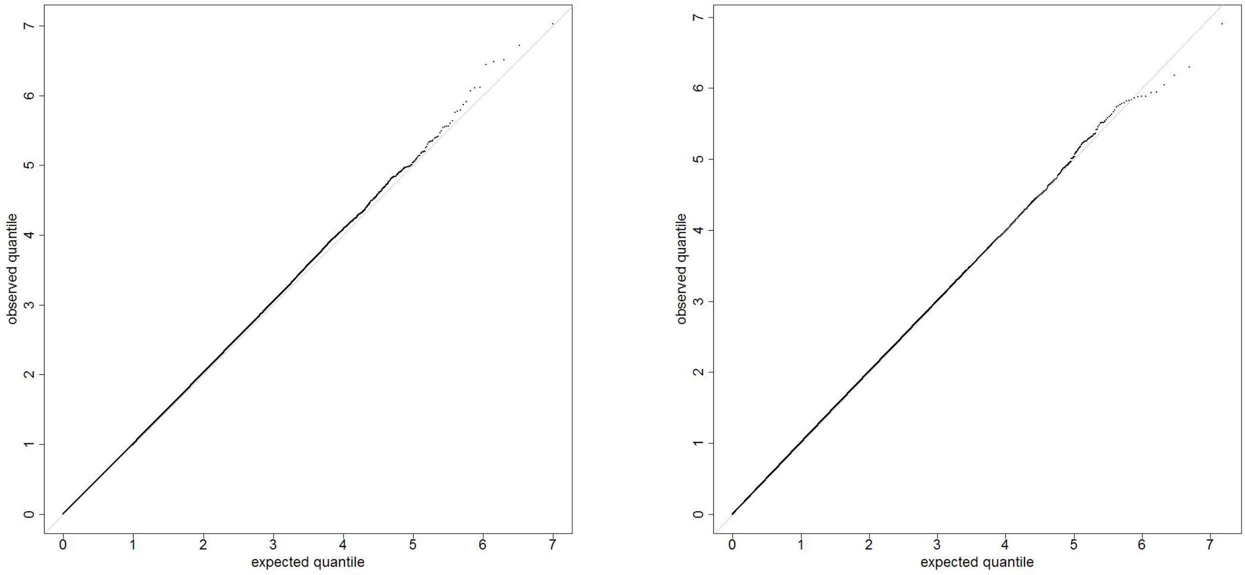
The number permutations is always an issue with permutation-based multiple testing. In data with a small sample size this number is necessarily be limited, but it quickly becomes very large already for moderate data sets. Although it would be best to use the collection of all possible permutations, this is often computationally not feasible, so a collection of randomly generated permutations is often used. Additional randomness is introduced in this way, which makes rejections and adjusted p-values random, especially if only few random permutations are used. The minimum number of permutations required depends on the method, the α-level, and on the presence of randomness in the permutations. The max$T$ method requires fewest permutations, and can work well with only $1/\alpha$ permutations, whatever the value of m, if p-values are continuous and all permutations can be enumerated. With random permutations a few more permutations are recommended to suppress randomness in the results, but a number of 1,000 permutations is usually quite sufficient at $\alpha = 0.05$, whatever m. The min$P$ method requires many more permutations. Because of the discreteness of the permutation p-values, the minimum observed p-value will be equal to the minimum possible p-value for most of the permuted data sets unless the number of permutations is very large, resulting in zero power for the method. For the min$P$ procedure, therefore, we recommend to use m/α permutations as an absolute minimum, but preferably many more. Such numbers of permutations are computationally prohibitive for typical values of m. Similar numbers of permutations are necessary for combinations of per-hypothesis permutation testing with Holm's or Hommel's procedure.

In the Van de Vijver data set we shuffled the survival status of the subjects 1,000 times, created a $4,919 \times 1,000$ matrix **P** of p-values, and performed the max$T$ procedure. The α-quantile of the distribution of the minimum p-values is found at 1.08×10^{-5}, which is remarkably close to the Bonferroni threshold of 1.02×10^{-5}, but still leads to 4 more rejections, for a total of 207. Stepping down by removing the rejected hypotheses leads to a slight relaxation of the threshold and one additional rejection, for a total of 208 rejections. In the Rosenwald data, the α-quantile of the minimum p-values evaluates to 7.86×10^{-6}, which is higher than the threshold of to 6.76×10^{-6} for Bonferroni, but does not lead to more rejections. Removal of the 4 rejected hypotheses does not alter the α-quantile, so the method stops at 4 rejections.

A gain in power for Westfall & Young's max$T$ method relative to Holm or Hommel can be achieved with Westfall & Young especially if strong positive correlations between p-values exist. The permutation method will adapt to the correlation structure found in the data, and does not have to take any worst case into account. A gain in power may also occur if the raw p-values are conservative. Permutation testing does not use the fact that p-values of true hypotheses are uniformly distributed, but adapts to the actual p-value distribution just as it adapts to the true correlation structure. Use of Westfall & Young does not require blind faith in the accuracy of the asymptotics underlying the raw p-values. Where methods that are not permutation-based become conservative or anti-conservative with the underlying raw p-values, Westfall & Young can even work with invalid and possibly anti-conservative p-values calculated from an incorrect model, and produce correct FWER control on the basis of such p-values. Although this sounds fabulous, it is sensible to be careful with this, however, since p-values from invalid models tend to be especially wild for probes for which the model fits badly, rather than for probes with an interesting effect. For this reason the power of a Westfall & Young max$T$ procedure based on such p-values is often disappointing. The min$P$ variant of the Westfall & Young procedure partially mends this by working on the per-probe permutation p-values instead of the raw p-values, guaratneeing a uniform distribution of the input p-values for the method.

In the case of the Van de Vijver and Rosenwald data, the asymptotic distribution of the p-values seems to hold reasonably well, as can be seen in Figure 3, a QQ-plot of the $_{10}$log of all the permuted p-values, although a slight tendency to anti-conservatism is noticeable in the Van de Vijver data. Such anti-conservatism would make the methods from previous sections anti-conervative, but has no effect on the Westfall & Young method.

Adjusted p-values can be easily calculated for the Westfall & Young procedure. They are always a multiple of $1/(k + 1)$ for random permutations, or of $1/k$ if all permutations can be enumerated [48]. For random permutations, letting α range over $1/(k + 1), 2/(k + 1), \ldots, 1$, and adjusting the α-quantile of the minimum p-value distribution accordingly, we can easily find the smallest of these α-levels that allows rejection of each hypothesis.



(a) Van de Vijver data

(b) Rosenwald data

Figure 3. QQ-plot of permuted p-values from the 1000 permuted data sets of Van de Vijver and Rosenwald at ${}_{10}$ log scale.

Software for the Westfall & Young procedures is available in the `multtest` package in `R`, and in SAS PROC MULTTEST.

2.5. Use of FWER control

Since the advent of genomics data, FWER as a criterion has been heavily criticised for being too conservative for genomics research. For many data sets, application of methods of FWER control result in very few rejected hypotheses or none at all, even when other analyses suggested the presence of some differential expression. This criticism of FWER stands at the basis of the development and popularity of the various FDR and FDP-based methods.

Indeed, FWER is a very rigorous and strict, and therefore conservative, criterion. It avoids type I errors at all cost, and as a consequence it introduces a large number of type II errors. The payback for this is that all hypotheses rejected by FWER controlling methods are individually reliable. FWER control implies $1 - \alpha$ confidence that each individual rejected hypothesis is correctly rejected. For many genomics experiments such great confidence is much more than necessary. If the experiment will be followed up by replication or proper validation before publication of the results, confidence that at least a substantial proportion of the findings is real is often sufficient to continue, and FWER-type confidence is not needed. What's more, at this stage the cost of type II errors is non-negligible, as missing out on an important finding can result in an expensive experiment wasted. More lenient criteria than FWER are in order for such experiments.

All this does not mean, however, that FWER has no place in genomics research. For the analysis of any experiments that are end-stage, not followed up by independent validation, such as the validation experiments themselves, merely saying that the proportion of true discoveries in the list is large is hardly sufficient. Such results have to be individually reliable, since they are likely to be taken out of the context of the list they were presented in. This individual reliability of results is precisely what FWER control guarantees. Since the power of validation experiments if typically large, and since the number of hypothesis tests done at this stage is limited, any conservativeness of FWER should negligible at this stage, especially if powerful methods such as Westfall & Young's are used.

3. Methods for FDR control

The seminal paper in which Benjamini & Hochberg [12] introduced the concept of FDR has changed thinking about multiple testing quite radically, showing that FWER control is not only way to do of multiple testing, and stimulating the field of multiple testing enormously.

Compared to FWER control, the subject of FDR control is relatively young. Much method-development is still ongoing, and some important questions are still partially open. This holds especially for the complicated situation of dependent p-values that is so important for applications in genomics research. In this paper, we leave aside the extensive literature on FDR control for independent p-values, and focus only on results that are known or believed to be valid under fairly general forms of dependence. We follow the same structure as for FWER-based methods, discussing methods that are generally valid, methods valid under the assumption that Simes inequality holds, and methods based on permutations. For FDR, unlike for FWER, the Simes-based method is the oldest and best known one, so we start there.

3.1. Benjamini & Hochberg

The Benjamini & Hochberg procedure [12] is a step-up procedure just like the Hochberg procedure, only with higher critical values. It compares each ordered p-value $p_{(i)}$ with the critical value $i\alpha/m$, finds the largest j such that $p_{(j)}$ is smaller than its corresponding critical value, and rejects the j hypotheses with the j smallest p-values. The Benjamini & Hochberg method is closely related to Simes' inequality (3), and the critical values of the Benjamini & Hochberg procedure are those of Simes' inequality with $m = m_0$.

It has been shown that this procedure controls FDR at level α if the PDSS assumption holds on the subset of true hypotheses [19]. This is the same assumption that underlies Simes inequality and Hochberg's and Hommel's methods. The Benjamini & Hochberg procedure has valid FDR control if the test statistics underlying the p-values are positively correlated for one-sided tests, or under more general dependence structures for two-sided tests. As discussed in Section 1.4, the validity of the procedure seems quite robust, at least for two-sided tests that are asymptotically normal. In fact, control of FDR under these assumptions is even at level $\pi_0\alpha$, where $\pi_0 = m_0/m$, rather than at level α.

The critical values of the Benjamini & Hochberg procedure are much larger than those of Hochberg or Hommel, so that many more rejections can often be made. In the Van de Vijver data, 1,340 hypotheses with p-values below 0.0136 are rejected at an FDR of 0.05, compared to 206 with p-values below 1.06×10^{-5} for Hochberg's method. In the Rosenwald data we reject 72 hypotheses with p-values below 4.86×10^{-4} with Benjamini & Hochberg, compared with 4 with p-values below 6.76×10^{-6} for Hochberg. Clearly, without changing the assumptions, relaxing the criterion from FWER to FDR can make a huge difference in terms of power.

A gain in power of Benjamini & Hochberg's method relative to Hochberg's, and in general of FDR-based versus FWER-based methods is most pronounced when many false hypotheses are present. This can be understood by comparing the FDR and FWER criteria. In FDR, the more hypotheses are rejected, the higher the denomimator of the false discovery proportion Q, and the less stringent the error criterion for the next rejection becomes.

The Benjamini & Hochberg method, like Bonferroni, controls its error rate at level $\pi_0\alpha$, rather than at α. This suggests the possibility an alternative, more powerful Benjamini & Hochberg procedure that uses critical values $i\alpha/(\hat{\pi}_0 m)$ rather than $i\alpha/m$ if a good estimate $\hat{\pi}_0$ of π_0 would be available. Such a procedure might have an FDR closer to the chosen level α, and would be even more powerful than the original procedure if many hypotheses were false. Such procedures are called *adaptive* procedures, and many have been proposed based on various estimates of π_0 [49]. A problem with the adaptive approach, however, is that estimates of π_0 can have high variance, especially if p-values are strongly correlated. Naive plug-in procedures, in which this variance is not taken into account, will therefore generally not have FDR control, especially if $\pi_0 \approx 1$. More sophisticated methods are needed that do take the estimation error of π_0 into account. One such procedure, by Benjamini, Krieger and Yekutieli [50], adjusts the α-level slightly from α to $\alpha^* = \alpha/(1 + \alpha)$ to adjust for the additional variance from estimation of π_0. This procedure estimates π_0 by first performing an initial Benjamini & Hochberg procedure at the slightly reduced level α^*, estimating π_0 by $\hat{\pi}_0 = (m - R_0)/m$, where R_0 is the number of rejections obtained in this first step. In the second and final step, a subsequent Benjamini & Hochberg procedure is done at level $\alpha^*/\hat{\pi}_0$. Note that, unlike simpler plug-in procedures, this latter procedure is not guaranteed to give more rejections than the regular, non-adaptive Benjamini & Hochberg procedure, since $\alpha^*/\hat{\pi}_0$ may be smaller than α. This reflects the additional risk incurred in estimating π_0. The adaptive procedure estimates $\hat{\pi}_0 = 0.73$ for the Van de Vijver data, resulting in 1,468 rejections, compared to 1,340 for the non-adaptive procedure. In the Rosenwald data the same procedure finds a disappointing $\hat{\pi}_0 = 0.99$, so that the critical value for the second stage is increased rather than decreased. A number of 69 hypotheses are rejected, compared to 72 for the non-adaptive Benjamini & Hochberg procedure. FDR control for the adaptive Benjamini, Krieger and Yekutieli procedure has only yet been proven under independence, although simulations suggest FDR control under positive dependence as well [50, 23, 51]. In any case, adaptive procedures are expected to have increased power over the ordinary Benjamini & Hochberg procedure only of the proportion π_0 of true null hypotheses is substantially smaller than 1. If π_0 is near 1, the power of such procedures is often worse. From a practical perspective, sacrificing power for the case that π_0 is near 1 in favor of power for small values of π_0 is seldom desirable: it increases the risk of not getting any rejections for poor data sets, while increasing the number of rejections in data sets in which there are already many rejections.

In Section 1.3 we argued that FDR control does not necessarily imply per comparison type I error control for individual hypotheses, and that procedures may sometimes reject hypotheses with p-values above α. The Benjamini & Hochberg method never does this, but adaptive variants might.

As a side note, we remark that adaptive control is not unique to FDR, and plug-in Bonferroni methods have also been suggested [52]. Just like for plug-in FDR, however, no proof of FWER control for such methods is available except under strong assumptions on the dependence structure of p-values.

Adjusted p-values for the procedure of Benjamini & Hochberg can be calculated using Algorithm 1. Adjusted p-values for FDR are sometimes referred to as q-values, but use of this term remains mostly connected to Storey's methods (Section 4.1). Some care must be applied when interpreting adjusted p-values based on FDR control, however, as we'll discuss in Section 3.4.

3.2. FDR control under general dependence

The equivalent to the Benjamini & Hochberg procedure that is valid even when the conditions for Simes' inequality does not hold is the procedure of Benjamini & Yekutieli [19]. This procedure is linked to Hommel's variant (2) of the Simes inequality in the same way that the procedure of Benjamini & Hochberg is linked with Simes inequality (3) itself. It is a step-up procedure that compares each ordered p-value $p_{(i)}$ with the critical value $i\alpha/(m\sum_{j=1}^{m} 1/j)$, finds the largest j such that $p_{(j)}$ is smaller than its corresponding critical value, and rejects the j hypotheses with the j smallest p-values. Like Hommel's inequality relative to Simes, the Benjamini & Yekutieli procedure is more conservative than the Benjamini & Hochberg procedure by a factor $\sum_{j=1}^{m} 1/j$. Like Hommel's inequality, the Benjamini & Yekutieli procedure is valid under any dependence structure of the p-values. Adjusted p-values for the Benjamini & Yekutieli method can be calculated using Algorithm 1.

The method of Benjamini & Hochberg is guaranteed to reject at least as much as Hochberg's procedure, which uses the same assumptions but controls FWER rather than FDR. The same does not hold for the method of Benjamini & Yekutieli relative to Holm's method, which is the standard method for FWER control under any dependence of the p-values. We see this immediately when we apply the Benjamini & Yekutieli procedure on the example data sets. In the Rosenwald data, where Holm's method rejected 4 hypotheses and Benjamini & Hochberg rejected 72, the procedure of Benjamini & Yekutieli, which effectively performs the Benjamini & Hochberg procedure at a level $\alpha/(\sum_{j=1}^{m} 1/j)$, which evaluates to 5.27×10^{-3}, allows no rejections at all. Comparing critical values, we see that the first $\log(m)$ critical values of Holm's method are larger than the corresponding critical values of Benjamini & Yekutieli. Therefore, if the expected number of false hypotheses is very small, Holm's method may be superior in terms of power to Benjamini & Yekutieli, and a better choice for FDR control. For less than extremely sparse data, however, we can expect Benjamini & Yekutieli to be more powerful than Holm. In the Van de Vijver data, where m is smaller and there are more false hypotheses, Benjamini & Yekutieli do reject substantially more than Holm, namely 614 hypotheses against 206 for Holm.

Alternatives to the Benjamini & Yekutieli method have been formulated by Sarkar [18] and by Blanchard and Roquain [53]. The latter authors proved that any step-up method with critical values of the form

$$c_i = \frac{\alpha}{m}\sum_{j=1}^{i} j f_j, \tag{5}$$

for non-negative constants $f_1, \ldots, f_m$ such that $\sum_{j=1}^{m} f_j = 1$, has control of FDR for any dependence struture of the p-values. Taking $f_j = 1/(j\sum_{k=1}^{m} 1/k)$ retrieves the Benjamini & Yekutieli critical values. Taking $f_j = 1/m$ retrieves the critical values proposed by Sarkar [18], given by

$$c_i = \frac{i(i+1)\alpha}{2m^2}.$$

Sarkar's method rejects 454 hypotheses in the Van de Vijver data, which is less than the method of Benjamini & Yekutieli. In the Rosenwald data, Sarkar's method, like Benjamini & Yekutieli, gives no rejections. A whole range of other FDR-controlling procedures also becomes possible in this way, parameterized by any chosen values of $f_1, \ldots, f_m$. As a motivation for choosing these values, it is helpful to realize that high values of f_j for some j make it relatively more likely that exactly j hypotheses will be rejected by the procedure [53]. From this, it is clear that Sarkar's method, even more than the method of Benjamini & Yekutieli, is focused on the situation that many hypotheses are false. No choice of $f_1, \ldots, f_m$ leads to an FDR controlling method that always rejects at least as much as Holm's method. As far as is currently known, therefore, also Holm's method remains admissible for controlling

FDR under general dependence, and it has the added boon of also controlling FWER. Still, Holm's method is only expected to be superior to Benjamini & Yekutieli in the situation that the number of false hypotheses is at most of the order of magnitude of $\log(m)$, so that $\pi_0 \approx 1$. In all other situations, the method of Benjamini & Yekutieli is a better choice.

3.3. FDR control by resampling

Several authors have worked on FDR control by permutation or by other types of resampling such as the bootstrap. However, an FDR controlling method with the power, simplicity and reliability of the method of Westfall & Young (Section 2.4) has not yet been found.

Research in this area is ongoing. The subject was pioneered by Yekutieli and Benjamini [54] who suggested a permutation-based procedure but without full proof of FDR control. Romano, Shaikh and Wolf [55], building on earlier work by Troendle [56] that had restrictive parametric assumptions, proposed a method use the bootstrap instead of permutations to control FDR asymptotically. Ge, Sealfon and Speed [57] proposed three different FDR-controlling methods, one of which has proven finite-sample FDR control. Its assumptions are more restrictive than those of the familywise error controlling method of Westfall & Young, but the method is still only marginally more powerful than that method, rejecting 209 hypotheses in the Van de Vijver data, one more that Westfall & Young, and 4 in the Rosenwald data, like Westfall & Young. None of the permutation FDR methods comes with user-friendly software.

3.4. Use of FDR control

As we have seen from the example data, FDR control is usually much less conservative than FWER control. Control of FDR, since that criterion is concerned with the proportion of type I errors among the selected set, is more suitable for exploratory genomics experiments than FWER control. FDR control methods do a very good job in selecting a set of hypotheses that is promising, in the sense that we can expect a large proportion of the ensuing validation experiments to be successful. As a consequence, FDR has effectively become the standard for multiple testing in genomics. Nevertheless, FDR control has been criticised [58, 59, 60, 61], sometimes heavily [14]. It is helpful to review some of this criticism in order to understand the properties and the limitations of FDR control better. Two main points of criticism concern the nature of FDR as an average.

In the first place, FDR is the expected value of FDP, which is a variable quantity because the rejected set $\mathcal{R}$ is random. It has been pointed out, however, that the actual value of FDP realized by FDR-controlling procedures can be quite variable, especially when p-values are dependent [59]. Sets $\mathcal{R}$ found by a method that controls FDR at α often have an FDP that is much larger than α, or one that is much smaller than α. The realized FDP for a method controlling FDR at 0.05 can, for example, be greater than 0.29 more than 10% of the time [58]. The variability of FDP around FDR is not taken into account in FDR control methods, and this variability is not quantified. Users of FDR must be aware that control of FDR at α only controls FDP in expectation, and that the actual proportion of false discoveries in the rejected set can often be substantially larger than α. FDR control is a property of the procedure leading to a rejected set, not of the rejected set itself.

Secondly, as we have noted in Section 1.3, FDR lacks the *subsetting property* that FWER does have. If a procedure controls FDR, the sets $\mathcal{R}$ generated have, on average, a false discovery proportion of maximally α. This property says something about the set $\mathcal{R}$, but the does not translate to subsets of $\mathcal{R}$ or to specific individual hypotheses that are elements of $\mathcal{R}$ [14]. Subsets may have much higher false discovery proportions than the complete set, and, since $\mathcal{R}$ is likely to contain a few false positives, each individual hypothesis in $\mathcal{R}$ may be such a false positive. In any case, the fact that $\mathcal{R}$ resulted from an FDR controlling procedure does not implicate any properties for subsets of $\mathcal{R}$. This lack of a subsetting property has several consequences that have to be taken into account when working with the results of FDR controlling procedures.

One consequence that has frequently been mentioned is the possibility of 'cheating' with FDR [14]. This cheating can be done as follows. If a researcher desires to reject some hypotheses using FDR, he or she can greatly increase the chances of doing so by testing these hypotheses together with a number of additional hypotheses which are known to be false, and against which he or she has good power. The additional guaranteed rejections alleviate the critical values for the hypotheses of interest, and make rejection of these hypotheses more likely. The catch of this approach is that the resulting FDR statement is about the rejected set including the added hypotheses, and that no FDR statement may, in fact, be made about the subset of the rejected set that excludes the added hypotheses. The cheating as described above is blatant, of course, and would hardly occur in this way. More often, however, inadvertent cheating of the same type occurs, for example when a list of rejected hypotheses is considered but the obvious, and therefore uninteresting, results are discarded or ignored, when an individual rejected hypothesis is singled out, or when subsets of the rejected hypotheses are considered for biological reasons. If hypotheses are very heterogeneous (e.g. Gene Ontology terms rather than probes) it is difficult not to look at subsets when interpreting the results on an analysis [61]. For correct interpretation of FDR control results, and to prevent inadvertent cheating, it is important to keep the rejected set complete.

A second consequence of FDR's lack of the subsets property relates to the interpretation of adjusted p-values. The adjusted p-value, being the largest α-level at which the hypothesis is rejected by the multiple testing procedure, is usually interpreted as a property of the hypothesis itself. For FWER-adjusted p-values this is warranted, as FWER control for the rejected set implies FWER control for each individual hypothesis within the set. FDR control is different, however: because that is a property of the whole set only, not of individual hypotheses within the set, the adjusted p-value similarly is a property of the whole rejected set, not of any individual hypothesis within that set. It is, therefore, hazardous to interpret such adjusted p-values as properties of individual hypotheses. To see this, consider a hypothesis with an FDR-adjusted p-value just below 0.05. If this hypothesis was the only hypothesis rejected at this α-level, we can be quite confident that this hypothesis is not a type I error. If, on the other hand, 20 or more hypotheses were rejected at the same level, the same adjusted p-value does not allow such a clear conclusion about that specific hypothesis. In the extreme, it could be that the top 19 hypotheses are absolutely certain rejections, and that the 20th hypothesis, even though a type I error with high probability, was added just because the FDR level of 0.05 left enough room for an extra rejection. Clearly, because of the lack of the subsetting property in FDR, interpreting FDR-adjusted p-values as properties of single hypotheses is problematic. An FDR-adjusted p-value is a property of a rejected set, not of an individual hypothesis. Interpreting it otherwise can be seen as a special case of "cheating with FDR".

An unrelated property has been shown for FDR can be very advantageous in incremental designs in which data are being gathered over a long period and repeatedly analyzed ad interim. FWER controlling methods must perform all interim analyses and the final analysis at a reduced level to correct for such multiple looks. With FDR, the need for such corrections all but disappears [62]. Allowing multiple looks, and, conversely, allowing more data to be added at a later stage after a first analysis, is a very valuable property in exploratory settings.

FDR controlling methods are most useful in exploratory situations in which a promising set of hypotheses must be found, and when we are content to use the complete set of hypotheses with top p-values, or something very close to that set. In such situations these methods are very powerful. If FDR control is used in a final reported analysis, the result of this analysis is the full rejected set, and researchers should be careful when making statements on subsets or individual hypotheses within that set.

4. Methods for FDP estimation

Next to methods for FDR control there are also methods that approach FDP from an estimation perspective. It pays to be precise about the quantities that are to be estimated, so we will discuss the difference between FDP estimation and FDR estimation briefly before turning to specific methods.

The purpose of an exploratory genomics experiment is to come up with a set $\mathcal{R}$ of discoveries, which are promising but might still require validation before they can be publicised. A measure of quality of such a chosen set $\mathcal{R}$ is the proportion of type I errors, i.e. the proportion of true null hypotheses present in this specific set, which is given by

$$q(\mathcal{R}) = \frac{\#(\mathcal{R} \cap \mathcal{T})}{\#\mathcal{R}}.$$

For any given set $\mathcal{R}$ this quantity is fixed but unknown. The purpose of FDP estimation methods is to come up with an estimate $\hat{q}(\mathcal{R})$ for the specific set $\mathcal{R}$ that is eventually chosen, and possibly to use the estimates for several competing sets in order to come up with a final choice for $\mathcal{R}$.

In contrast to FDP estimation, FDR estimation aims to estimate FDP on average over the distribution of rejected sets $\mathcal{R}$ induced by the test procedure, rather than for the specific chosen set. FDP is a property of the rejected set; FDR is a property of the test procedure [60]. The value of FDP depends only on truth and falsehood status of hypotheses; the value of FDR also depends on many other (nuisance) parameters, and on the sample size, because the distribution of $\mathcal{R}$ depends on such parameters.

The potential for FDP estimation, as opposed to FDR or FWER control, is that it allows researchers to approach multiple testing in a different way. Methods that control an error rate require the user to specify the error rate to be controlled and the level at which it is to be controlled before seeing the data. Next, the multiple testing procedure selects the rejected set $\mathcal{R}$ as a function of the data and the chosen error rate, without any further input from the user. With FDP estimation, the role of the user and the multiple testing procedure can be reversed. If reliable estimates of $q(\mathcal{R})$ are available for every set $\mathcal{R}$ of potential interest, the user can review these sets and estimates and select the most promising one, and the role of the multiple testing procedure can be to inform the user of the likely FDP in the set by providing an estimate of $q(\mathcal{R})$. This reverses the traditional roles: the user chooses the rejected set, and the multiple testing method specifies the error rate.

Two issues, however, must be taken into account by FDP estimation methods in order to make this approach viable. The first is selection. Finding an estimate of $q(\mathcal{R})$ for a fixed set $\mathcal{R}$ is relatively easy, but the rejection set $\mathcal{R}$ of interest is typically chosen on the basis of the same data that are also for FDP estimation, and this set will often probably be chosen because it has a small estimate $\hat{q}(\mathcal{R})$. The value of $\hat{q}(\mathcal{R})$ for the post hoc selected $\mathcal{R}$ is therefore very likely to be underestimated. Honest estimation of $q(\mathcal{R})$ should protect against this estimation bias or correct for it. The second difficulty is to provide an assessment of the uncertainty in an estimate $\hat{q}(\mathcal{R})$. Giving a point estimate of FDP without any assessment of its variability is not very informative. The variance of the estimate, however, is crucially influenced by the dependence structure of the p-values used for the estimation. Any assessment of this variability is again subject to the above-mentioned problem of selection bias due to the selection of $\mathcal{R}$.

We focus in this section on methods that explicitly concentrate on FDP rather than FDR estimation, and which do not only provide a point estimate, but also assess its variability by providing confidence statements for FDP. As with FWER and FDR control we discuss Simes-based and worst-case methods (Section 4.3) as well as permutation methods (4.4). Before going into such methods, we briefly review some older estimation approaches, which take an empirical Bayes perspective: Storey's (Section 4.1) and Efron's (Section 4.2). Both methods provide point estimates only.

4.1. Storey's approach and SAM

The first one to suggest an estimation perspective for FDR was Storey [63], who was motivated by the empirical Bayesian view of FDR [13]. Storey considered only sets of rejected hypotheses of the form $\mathcal{R} = \{H_i : p_i \leq t\}$ for some constant t. In such a rejected set, the expected number of true hypotheses to be rejected is $m_0 t$, because p-values of true null hypothesis follow a uniform distribution. This number can be estimated by substituting an estimate $\hat{m}_0$ for m_0. Storey suggests

$$\hat{m}_0 = \frac{\#\{p_i > \lambda\} + 1}{1 - \lambda}, \tag{6}$$

where $0 \leq \lambda < 1$ is some constant, typically taken as $1/2$. To understand this estimator, remark that a proportion $1 - \lambda$ of p-values of true hypotheses is expected to be above λ, but a smaller proportion of p-values of false hypotheses, so that $\mathrm{E}(\#\{p_i > \lambda\}) \geq m_0(1 - \lambda)$. Consequently, $\mathrm{E}(\hat{m}_0) \geq m_0$, making $\hat{m}_0$ a conservative estimate of m_0. The addition of 1 to the numerator makes sure that $\hat{m}_0^{-1}$ is always defined. Using this estimate, Storey writes $\mathrm{FDR} \approx m_0 t/\#\mathcal{R}$, so that

$$\hat{q}(\mathcal{R}) \;=\; \frac{\hat{m}_0 t}{\#\mathcal{R}} \;=\; \frac{t(\#\{p_i > \lambda\} + 1)}{(1 - \lambda)\#\{p_i \leq t\}}. \tag{7}$$

This estimate was presented by Storey as an FDR estimate, but it might be better viewed as an FDP estimate.

Storey's estimate was derived under the assumption of independence among p-values. At first sight, it appears hardly affected by dependence among p-values. The expected number of true null hypotheses with p-value smaller than t is $m_0 t$ whatever the joint distribution of these p-values. Dependence, however, does play a large role in the variability of the estimate, since the number of true null hypotheses with p-values below t can have high variance especially if p-values are positively correlated, and this may affect the performance of the estimate. A proof of conservative consistency for the estimate is available [64], which says that $\lim_{m\to\infty} \hat{q}(\mathcal{R}) \geq q(\mathcal{R})$. However, this property has been shown to hold only under the assumption of independent p-values or under a form of weak dependence that allows only local correlations, and may fail otherwise [65]. This weak dependence assumption seems too weak to be useful for genomics data, except, perhaps, genome-wide association studies, which have mostly local dependence between markers. Under more realistic dependence, Storey's estimates are very variable and can underestimate the true FDP by a wide margin [66, 67, 68, 23, 55, 51]. A highly variable estimate does not have to be a problem, as long as we can assess this variability. Some important work has been done in finding the distribution of the estimator [67, 65], but no definite methodology has emerged. Storey's estimator can have very large variance, large positive skewness, and substantial bias.

Proof that post hoc choice of the threshold t, which defines $\mathcal{R}$, is allowed, is only available under the assumption of independent p-values. In other situations, it is unclear how such a choice biases the final estimate.

Storey's estimate is sometimes used as a way to control FDR, rather than as a way to estimate FDP. This is done by selecting the highest value of t such that the estimate (7) is below α. This actually links quite closely to the Benjamini & Hochberg procedure. If instead of using (6) we estimate m_0 conservatively by m, then we have $\hat{q}(\mathcal{R}) = mt/(\#\{p_i \leq t\})$. Finding the largest t such that this estimate is below α is exactly achieved by the

Benjamini & Hochberg procedure. Similarly, if Storey's approach with the estimate of π_0 is used, this results in an adaptive Benjamini & Hochberg procedure, which has the problems associated with naive plug-in procedures as described in Section 3.1. There is no proof of FDR control for Storey's procedure. In general we should not expect to succeed in keeping the true value of a quantity below α by keeping an estimate of that quantity below α. FDR control based on FDP or FDR point estimation is less reliable than true FDR control.

The method known as Significance Analysis of Microarrays (SAM) can be viewed as a permutation based variant of Storey's method. It has been shown, both using theory [1], and by simulations [23] that SAM does not control or reliably estimate FDP or FDR, but instead estimates $\mathrm{E}(V)/\mathrm{E}(R)$. Since the latter quantity is easier to control, SAM tends to be much less stringent than FDR-controlling methods.

4.2. Efron's approach

Where Storey was merely motivated by the empirical Bayes approach to FDR, Efron [69] took this view much further. Rather than only assuming that truth of hypotheses was random, he suggested to make the additional assumption that the test statistics $T_1, \ldots, T_m$ corresponding to hypotheses $H_1, \ldots, H_m$ were drawn i.i.d. from a mixture distribution with density

$$f(t) = \pi_0 f_0(t) + (1 - \pi_0) f_1(t). \tag{8}$$

Here, π_0, again, is the prior probability that a hypothesis is true, $f_0(t)$ is the density of the test statistics of true hypotheses, and $f_1(t)$ the density of test statistics of false hypotheses. The value of π_0 and the shape of f_1 are unknown and must be estimated from the data. The form of f_0 is usually considered known, but may sometimes also be estimated.

The assumption (8) is a very powerful one. It allows all kinds of very specific statements to be made on the conditional probability that a certain hypothesis is true given what has been observed about this hypothesis. In particular, one can calculate $\mathrm{P}(H_i \text{ is true} \mid T_i \geq t)$, which yields FDR statements similar in spirit to Storey's, and

$$\mathrm{P}(H_i \text{ is true} \mid T_i = t),$$

which Efron calls the *local FDR*. This is a powerful FDR-type statement about an individual hypothesis.

At first sight, the assumption (8) hardly seems an assumption at all. In fact, plotting a histogram of the observed test statistics coming from a genomics experiment will always produce a distribution that looks very much like the one suggested by (8). On more close inspection, however, the assumption (8) is a very strong one, and one that is difficult to reconcile with genomics experiments. The crucial part of the assumption is that the test statistics are drawn *i.i.d.* from the distribution (8). We will discuss the identically distributed part first, before going into the independence assumption.

If the test statistics of all hypotheses are drawn from the same distribution (8), in particular every hypothesis has the same a priori probability π_0 of being true. By this assumption, the presence of many false hypotheses, through a low estimated π_0, may result in low local FDR estimates for all other hypotheses. This assumption, therefore, essentially legitimizes the "cheating with FDR" phenomenon, explained in Section 3.4. In FDR control, this phenomenon was a result of misinterpretation of FDR control results. Under the assumption (8), interpreting FDR results as pertaining to individual hypotheses, as local FDR does, becomes legitimate, because the inhomogeneity of hypotheses that is necessary for "cheating with FDR" is simply assumed not to exist. Relaxing the identical distribution assumption, e.g. by giving each hypothesis its own π_0, makes the model (8) unidentifiable without substantial external information.

The independence assumption on the test statistics drawn from (8) is essentially an assumption that genomic probes are independent of each other. Even though such an assumption is not warranted in genomic data, it has been argued that point estimates resulting from a model that assumes independence may still be useable in the presence of dependence. The same arguments come back here that played a role in Storey's method, so we only briefly reiterate them here. Dependence of test statistics may not have a very big impact on point estimates, it is a big determinant of variance of such point estimates [70]. Because of the reliance on the independence assumption, Efron's methods can only report the estimates themselves, but cannot assess their variability in the presence of dependence. In general, reporting a point estimate without any assessment of variability can too easily give a false sense of security. Recent variants of Efron's methods have attempted improve point estimates under dependence by replacing the assumption that test statistics of true hypotheses are drawn i.i.d. from a known f_0 to the assumption that they are drawn i.i.d. from another distribution $\tilde{f}_0$, to be estimated from the data [71, 72, 73]. Although this improves the accuracy of point estimates in case of dependence, it still does not allow the variability of these estimates to be assessed.

Correction for bias due to selection of the hypotheses with the best local FDR has, as far as we know, never been discussed in relation to Efron's methods.

4.3. FDP confidence

Working from a frequentist, not empirical Bayes perspective, Goeman and Solari [11] derived confidence statements for the FDP $q(\mathcal{R})$, instead of looking for point estimates. Confidence statements have the advantage over point estimates that they automatically take variability into account. More precisely, Goeman and Solari derive an upper confidence bound $\bar{q}_\alpha(\mathcal{R})$, and prove that a confidence statement holds of the form

$$\mathcal{P}(q(\mathcal{R}) \leq \bar{q}(\mathcal{R})) \geq 1 - \alpha. \tag{9}$$

In this equation, the set $\mathcal{R}$ can be any subset of the collection of hypotheses $\mathcal{H}$ of the users choice. Importantly, this confidence bound holds for every possible set $\mathcal{R}$ simultaneously, so that, by the properties of simultaneous confidence bounds, the validity of equation (9) is not compromised by multiple looks at the data, and the confidence statement holds for a post hoc selected set as well as for any other set. All these confidence statements can be alternatively presented as confidence statements on $v(\mathcal{R}) = \#(\mathcal{R} \cap \mathcal{T})$ by simply multiplying by the fixed quantity $\#\mathcal{R}$ before and after the first inequality sign in (9).

Since the method allows sets $\mathcal{R}$ of any form, not just sets consisting of the hypotheses with the best p-values, the method of Goeman and Solari accommodates the tendency of researchers to cherry-pick among the list of top genes coming from an experiment, composing a meaningful selection of hypotheses to take to the next stage of validation using a mixture of statistical and biological considerations.

Two variants of the method of Goeman and Solari are relevant for this overview: one that assumes that Simes' inequality holds, and one that holds for general dependence structures of p-values. The Simes-based method uses exactly the same combination of the closed testing procedure [42] and the Simes inequality that the methods of Hochberg and Hommel use. Only, instead of using this procedure only to derive FWER-type statements for the individual hypotheses, they obtain simultaneous confidence bounds of the form (9) using the same procedure. Since all confidence bounds are derived from a single application of the closed testing procedure, they depend on the same event for their coverage, making these confidence bounds simultaneous. Because the underlying method is the same, the assumption of PDSS of the true null hypotheses underlying the Simes-based method of Goeman and Solari is identical to the assumption underlying Hommel's and Hochberg's methods, and identical to that underlying the Benjamini & Hochberg procedure.

The result of the method is not a single rejected set, such as the methods from Sections 2 and 3, but rather $2^m - 1$ simultaneous confidence bounds, one for every possible subset $\mathcal{R}$ of $\mathcal{H}$. In the Van de Vijver data, taking $\mathcal{R} = \mathcal{H}$, we find that with 95% confidence there are at least 640 false null hypotheses among the 4919. The smallest set containing at least 640 false null hypotheses at this confidence level is the set of 837 hypotheses with smallest p-values. If we would reject this set, the FDP of our findings would be at most (837-640)/837 = 23.5%, with 95% confidence. The connection with Hommel's method becomes obvious if we take $\mathcal{R}$ to be the set of 209 hypotheses rejected by Hommel's method. This set is the largest set for which we find $\bar{q}(\mathcal{R}) = 0$, which coincides precisely with the FWER statement obtained from Hommel's methods, stating that with 95% confidence each of these 209 rejections is a correct one. In the Rosenwald data, with 95% confidence at least 14 out of the 38 hypotheses with best p-values are false, yielding an FDP confidence interval ranging from 0 to 63.2% for this set. So far we only considered sets $\mathcal{R}$ of a type consisting of the k hypotheses with best p-values. We may also, however, take some other set, perhaps partly chosen for biological reasons. For example, in the Rosenwald data we may select a set $\mathcal{R}$ consisting of the hypotheses with p-values ranked $\{2, 5, 6, 7, 8, 9\}$ and find that this set has 95% confidence of an FDP of at most 50%. Since all confidence statements obtained from the same data set are simultaneous, they remain valid even if the researcher reviews many possibilities and finally picks a result that stands out.

Comparing the above results to FDR control by the method of Benjamini & Hochberg, which is valid under the same assumptions as the Simes-closed testing combination used by Goeman and Solari, we see that the set of 1340 hypotheses with best p-values rejected in the van de Vijver data by Benjamini & Hochberg only gets an FDP upper confidence bound of 52.2%. This partly reflects the great variability of FDP: although on average sets selected by the Benjamini & Hochberg have an FDP of at most 5%, the variability of FDP around FDR is large, and the confidence interval for FDP for this particular set ranges from 0 to 52.2%. In the Rosenwald data, the set of 72 hypotheses selected by Benjamini & Hochberg gets a confidence interval for its FDP of 0 to 80.6%. For both data sets, there are smaller rejected sets with much better FDP bounds, as we have seen above. Because of mathematical similarities between the Benjamini & Hochberg procedure and the Simes-based procedure of Goeman and Solari, it can be shown that the set rejected by Benjamini & Hochberg always gets an upper confidence bound for FDP that is strictly less than 1.

If desired, point estimates can be calculated for FDP by simply taking a 'midpoint' of the confidence interval, i.e. using $\bar{q}(\mathcal{R})$ at $\alpha = 0.5$ as an estimate. Calculated by means of a simultaneous confidence interval, this point estimate is robust to selection of $\mathcal{R}$. In the van de Vijver data the set of top 837 hypotheses, which had 95% confidence of an FDP at most 23.5%, gets a point estimate for FDP of only 1.7%. Similarly, the top 38 hypotheses in the Rosenwald data get an FDP point estimate of 5.3%, with 95% confidence interval [0%, 63.2%]. The α-midpoint of the confidence interval is often much closer to 0 than the full confidence interval would seem to suggest. Point estimates, however, can augment the FDP confidence intervals, but are not sufficient by themselves as a basis for inference.

A variant of the FDP confidence bound method is based on Hommel's inequality (2), which reduces all critical values of the Simes inequality by a factor $\sum_{k=1}^{m} 1/k$. This method is valid for all dependence structures of the p-values. It relates to the Benjamini & Yekutieli method in the same way that the Simes-based method relates to the Benjamini & Hochberg method. In the Van de Vijver data, the confidence bounds arising from this method say with 95% confidence that 284 false null hypotheses are present among the 385 hypotheses with best p-values, with a point estimate of FDP of 2.3%. In the Rosenwald data, the same method returns the trivial confidence bound $\bar{q}(\mathcal{R}) = 1$ for every $\mathcal{R}$, which will happen for any data set in which the Benjamini & Yekutieli method yields no rejections. Just like the Benjamini & Yekutieli method, the Hommel-based method of calculating FDP confidence bounds can be quite conservative, and is sometimes less powerful than a simple application of Holm's method.

The FDP confidence method takes into account variability in the FDP estimate, by providing confidences intervals rather than a point estimates. It also takes into account post hoc selection of $\mathcal{R}$, by making the FDP confidence intervals simultaneous for all $\mathcal{R}$. These two properties make the method suitable for the reversal of roles described in the beginning of Section 4. After looking at the data, the user of this method can select the set $\mathcal{R}$ of rejected hypotheses freely, and be informed of the maximal proportion of false rejections made, at the chosen confidence level, when rejecting this set. On the basis of this assessment the user can revise the set, once or many times, to come up with a final set $\mathcal{R}$ with its FDP confidence bound $\bar{q}(\mathcal{R})$. The validity of the final FDP confidence interval $[0, \bar{q}(\mathcal{R})]$ is not compromised by the selection process.

FDP confidence bounds are defined for a fixed confidence level $1-\alpha$, and therefore do not easily admit the use of adjusted p-values. Analogues of adjusted p-values can be given [74], but since these do not take the form of a single value, they are less straightforward to use.

Methods for calculating FDP confidence bounds are available in the `R` package `cherry`.

4.4. Meinshausen's permutation method

Meinshausen [75] developed methods based on permutations for making statements of the form (9). His confidence bounds were initially only for rejected sets of the form $\mathcal{R} = \{H_i : p_i \leq t\}$, but they are simultaneous over all choices of t. Goeman and Solari [11] showed how to extend Meinshausen's method also to obtain simultaneous statements for general sets $\mathcal{R}$, without loss of power, by embedding Meinshausen's procedure in a closed testing procedure.

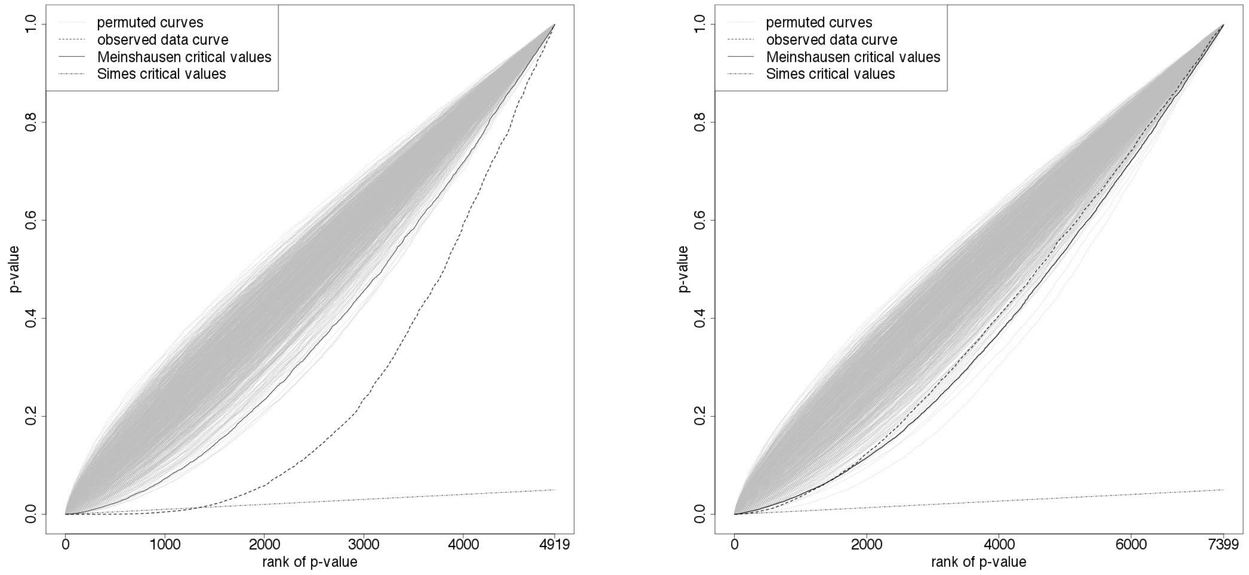
Meinshausen's method is based on a permutation-based variant of Simes' inequality, finding critical values $c_1, \ldots, c_m$ such that with probability at least $1-\alpha$, we have

$$q_{(i)} \geq c_i \qquad \text{for all } i = 1, \ldots, m_0, \tag{10}$$

where $q_{(1)} \leq \ldots \leq q_{(m_0)}$ are the m_0 ordered p-values of hypotheses corresponding to true null hypotheses. Rather than obtaining these critical values on theoretical grounds as Simes did, Meinhausen finds them by permutation arguments. Constructing a matrix $\mathcal{P}$ of permutation p-values as in Westfall & Young's max$T$ (Section 2.4), each permuted data set provides a ranked p-value curves such as displayed in Figure 1. The values of $c_1, \ldots, c_m$ are chosen in such a way that a proportion $1-\alpha$ of these permuted curves are everywhere larger than the curve of these critical values. Either directly, or through closed testing arguments, the critical values obtained can be used in the same way as the Simes critical values are used by Goeman and Solari to obtain confidence statements of the form (9).

The ranked p-value curves for 1000 permutations of the data, and the resulting critical values calculated from these curves by Meinshausen are illustrated in Figure 4. What is immediately visible is the large variability of the p-value curves around the diagonal line. This variability is larger than would be expected if the p-values were independent and illustrates the dependence of the p-values. Comparing the critical values of Meinshausen and Simes, the first few Simes critical values tend to be larger than those of Meinshausen, but the Meinshausen critical values are larger over the rest of the domain. As a consequence, Simes-based methods tend to give tighter confidence bounds for small rejected sets, but Meinshausen's method gives tighter bounds for larger rejected sets.

To use Meinshausen's method, valid permutations must be available, as described in Section 1.4, just like for Westfall and Young. Like the methods of Westfall and Young, Meinshausen's method is robust against p-values that are not marginally properly uniform. We did find, however, that Meinshausen's method can become anti-conservative if a limited number of randomly generated permutations is used, due to overfit of the critical value



(a) Van de Vijver data

(b) Rosenwald data

Figure 4. p-value curves calculated for observed and permuted data sets, with Simes' critical values and those calculated by Meinshausen on the bases of the permutations.

curve to the chosen permutations. We recommend to use a large number of permutations if not all permutations can be enumerated.

Applying Meinshausen's method on the Van de Vijver data, we find with 95% confidence that at least 1121 hypotheses out of 4919 are false, and that the smallest collection of hypotheses to contain at least 1121 false hypotheses with 95% confidence is the set of 2008 hypotheses with the best p-values. For these sets, Meinshausen's result is quite a bit stronger than the Simes-based result of Goeman and Solari, but this reverses for smaller sets: where Goeman and Solari reproduced the Hommel FWER result that the 209 hypotheses with best p-values contained no true hypotheses at 95% confidence, Meinshausen's method can make the same assertion only for the top 99 hypotheses. The same trend is visible in the Rosenwald data, where Meinshausen finds as many as 166 false hypotheses to be present among the 562 hypotheses with best p-values, but finds no individual hypotheses to be false with 95% confidence. Comparisons of Meinshausen's method with FDP confidence bounds based on the more conservative Hommel's inequality all fall out clearly in favor of Meinshausen's method. Point estimates of FDP can be calculated for any sets in the same way as for the methods of Section 4.3, but unbiased point estimated are also available [76].

Meinshausen's method makes the same type of simultaneous confidence statements as Goeman and Solari's methods in Section 4.3, and can be used in the same way, letting the user repeatedly choose the rejected set $\mathcal{R}$, and letting the method advise about the proportion of false positives present in that set. Software for applying Meinshausen's method and variants is available in the `howmany` and `cherry` packages in `R`.

4.5. Use of FDP estimation

FDP estimation methods are tailored to the exploratory setting. They allow the researcher to obtain an estimate or a confidence statement for any set $\mathcal{R}$ that he or she may choose to reject, and the option to pick one or several sets that are biologically meaningful as well as statistically reliable, without paying a price for this cherry-picking in terms of additional false positives. This seems the epitome of exploratory research, empowering researchers and giving them tools to do the selection of interesting findings, rather than taking the selection process out of their hands. FDP statements also relate directly to the actual rejected set, rather than only indirectly through a property of the procedure. Just like with FWER and FDR control methods, however, care must be employed when using these methods. Methods that can also provide confidence bounds for FDP are preferable to methods that merely give point estimates, so we will concentrate on the FDP confidence methods from Sections 4.3 and 4.4 here.

It should be realized that all FDP estimates relate to sets, and that FDP statements about sets do not say anything directly about individual hypotheses that are part of these sets. This is the same cautionary statement as we made in Section 3.4 about FDR. If a set of 100 hypotheses has an FDP of maximally 0.01 with 95% confidence, then each individual hypothesis among these 100 may or may not be a likely candidate for a false positive. Unlike with FDR, however, this problem is alleviated by the fact that FDP confidence methods do simultaneously provide confidence bounds for all subsets of the set $\mathcal{R}$. These simultaneous subset statements at least partially solve the cheating problem (Section 3.4), because a valid confidence statement is always available for the set of hypotheses of real interest. The methods also allow statements to be made for individual hypotheses, by obtaining confidence statements for subsets of $\mathcal{R}$ of size 1. To completely avoid problems of overinterpretation of FDP statements about sets, however, ideally all $2^m - 1$ simultaneous confidence statements should be reported. This is impossible, of course, and the sets for which the confidence bounds are calculated and reported should be chosen carefully.

Because of the identity between the underlying methods, FDP confidence statements actually come for free with the use of Hochberg's and Hommel's methods. Using Hommel's FWER control and FDP confidence together, the FWER-controlled rejections are augmented by weaker statements for the set of hypotheses that just did not make the FWER threshold, claiming that at least a proportion of these hypotheses is false. These may be forwarded to further validation experiments if the FDP confidence bound for such a set is promising enough.

FDP confidence methods are most fruitfully used in exploratory settings if much freedom is desired in picking and choosing the set of rejected hypotheses of interest, or if more specific and more confident FDP statements are required than those provided by FDR control.

5. Reducing the number of tests

So far, we have assumed that the collection of hypotheses to be tested was fixed a priori. In practical testing problems this does not have to be the case. Even if tens of thousands of probes or more have been measured on a genomic chip, this does not mean that all of these data as such have to be fed into the testing procedure. In fact, there are large gains to be had by either selecting a more limited number of hypotheses to be tested, or by aggregating hypotheses in order to test them jointly. In practice these gains can be much greater than the gain obtained by changing from one multiple testing procedure to another more powerful one. We discuss selection and aggregation in turn.

5.1. Selection

The more hypotheses, the bigger the multiple testing problem. This holds immediately for FWER control methods and FDP confidence methods, which are always more powerful if fewer hypotheses are being tested. By this we mean that discarding some hypotheses before performing the testing procedure always leads to more rejections or tighter confidence bounds among the remaining hypotheses than discarding the same hypotheses after performing the same procedure. The same is not always true for FDR-controlling methods, as we have seen in Section 3.4, but also for those methods removing hypotheses prior to testing tends to result in more power, especially if the discarded hypotheses are predominantly either true hypotheses or false hypotheses for which the test has low power. Anyway, with FDR control, any discarding needs to be done before testing, since discarding after testing risks loss of FDR control. Of course, discarding of hypotheses brings the risk of discarding hypotheses that would have been highly significant if not discarded. Discarded hypotheses should therefore be chosen either because they are uninteresting, or because they have low power.

Uninteresting hypotheses are hypotheses whose rejection the researcher would not follow up on even if they came out with the smallest p-value. Surprisingly many of such uninteresting hypotheses are tested in genomics experiments. Many researchers, for example, are only interested in annotated probes, i.e. probes for which the name of the associated gene is known. Tests of non-annotated probes typically also have low power, because they are often low-expressed, and therefore seldom show up on the top of the list of significant genes. If they do, they are often either ignored or thrown out after the analysis. Non-annotated probes make up over 35% of the probes in the Van de Vijver data. Using Holm's method as an example, out of 206 hypotheses rejected without selection, 158 are annotated. Removing non-annotated probes before testing, the number of annotated probes rejected increases to 176.

Discarding hypotheses with low power before testing can also benefit the overall power of the multiple testing procedure. If there are hypotheses that are very unlikely to give a small enough p-value to be rejected, the only effect of including them is to worsen the critical values of more promising hypotheses. Deciding which hypotheses have insufficient power is often done on the basis of the data, but only certain aspects of the data may be used without compromising type I error control. As a rule of thumb the selection criterion must be independent of the p-values, but precisely what can or should be used depends on the specifics of model and data [77, 78, 79, 80]. In gene expression data, a viable and powerful selection is to discard probes with low mean or low variance under the null hypothesis, since both of these are indicators of low biological variation in gene expression. In the Rosenwald data, removing the 50% probes with low mean expression, we get 6 rejections with Holm's method rather than 4. Compared to a priori discarding of uninteresting probes, selection on the basis of presumed power is more risky: good results may be discarded with the bad. We see an example of this in the Rosenwald data if we select on high variance of expression instead of high mean: this leaves us with only 3 rejected probes. Researchers concerned about this risk may want to consider weighted multiple testing as an alternative to selection, down-weighting rather than discarding hypotheses with low power [77, 78].

It is interesting to remark that the two selection criteria of interestingness and power often at least partially coincide. In the examples above, on the one hand non-annotated probes are often low expressed, and therefore typically have low power. On the other hand, probes with low mean or variance of expression are unlikely to have a large magnitude of differential expression, even if they would have a very low p-value. Such a significant probe with a small effect size is less trustworthy and is often found less interesting than one with a large effect size.

5.2. Aggregation

A different way of reducing the size of the testing problem is to aggregate the data to a lower resolution. For example, rather than testing every probe, aggregated tests can be performed at the gene level or at the level of a chromosomal region. Such aggregation is especially beneficial if the aggregated level is the level of primary biological interest. Not only does aggregation reduce the size of the testing problem, aggregated tests also tend to be more powerful than non-aggregated ones.

The choice of an aggregated test should be determined by the type of data, the level of aggregation, and domain knowledge. In gene expression data, for example, we would expect different probes of the same gene to show approximately the same amount and direction of differential expression. With this in mind, a suitable aggregate test at the gene level would just calculate the gene's expression as an average of each probe's expression, and use this as a basis for testing. When aggregating to higher levels than the gene, or when the possibility of differential splicing leads us to expect highly heterogeneous differential expression of probes, then more sophisticated aggregated tests may be preferable [38, 39]. In all cases, the formal null hypothesis of an aggregated test is that the hypotheses of all of the probes that have been aggregated are true. In the Van de Vijver data the 3179 annotated probes represent 3043 genes, so there is not much to be aggregated. Still, taking Holm's method as an example again, we see that the 176 annotated probes rejected in Section 5.1 correspond to 170 genes. Aggregating to the gene level before testing, using simple averaging of probes, improves this to 173 rejected genes.

If not only the aggregated level is of interest, but also lower or higher levels of resolution, then it is possible to use hierarchical multiple testing methods that test more than one level of resolution simultaneously. Such methods start testing at the most aggregated level, continuing to a more detailed resolution only at those locations for which significant results have been found. Hierarchical methods are available for both FDR and FWER control [81, 82, 83, 84]. This type of methods can also be used in the situation that multiple hypotheses are tested about the each probe, by first rejecting the overal null hypothesis for each probe that all these hypotheses are true, before going down to the detailed tests of the individual hypotheses for that probe.

6. Discussion and conclusion

A typical genomics experiment involves many thousands of hypothesis tests. It is unavoidable, therefore that both the researcher him- or herself, and the reader of the resulting publication, only sees a very small selection of the results of these tests. Human intuition is not capable of quantifying the effect of this selection: formal methods are needed to correct for it. This is what multiple testing methods are for.

From this perspective, there can be no reason not to correct for multiple testing in a genomics experiment. Even if a user is not prepared to let the results of a multiple testing method dictate his or her actions, the results of a multiple testing method carry information about the effect of selection on the reliability of any conclusions drawn. The least such a user can do is to estimate the number of false positives in the selected set, as the methods of Section 4 allow to do.

Multiple testing methods traditionally only concentrate on the effect of selection on p-values and on rejection of hypotheses. Selection, however, also biases effect size estimates. The estimated fold change of differential expression of the probe with highest expression is very likely to be overestimated, and is therefore subject to a regression to the mean effect upon replication of the experiment. This is the famous *winner's curse*, which has been much discussed in the context of genomewide association experiments, where it shows up in the odds ratios of top SNPs [85].

One other aspect that is universally ignored by multiple testing methods is the haphazard trying of different models and analysis techniques that is typical of actual data analysis. Models are fitted with and without certain covariates; several tests may be tried. In the end, it is more likely that a model with many significant results is selected for the final analysis than one with few. Such model tuning may severely bias the p-values downward, and increase the likelihood that type I errors appear. Since this process is highly unstructured, it is impossible to formally correct for it. The resulting bias can be prevented by writing a complete analysis plan before data collection, and adhering strictly to it, but this is something few researchers in genomics experiments will be prepared to do. It can also be prevented by good and independent validation experiments.

It cannot be stressed enough that if any multiple testing is applied anywhere, it must be in the independent validation experiments. The results of validation experiments must be able to hold their own, individually, and without further evidence to back them up. Familywise error control is the norm here. If a genomics experiment is not followed up by independent biological validation, it must be very strict in its multiple testing correction, since the experiment essentially doubles as its own validation.

When choosing a multiple testing correction method, two important questions need to be asked. First, what assumptions can be made about the joint distribution of the p-values observed in the experiment? Second, what type of multiple testing method is needed? We will finish this tutorial by briefly reiterating some of the considerations that can play a role when answering both these questions.

Regarding the assumptions on the p-value distribution, it is important to realize that p-values in genomics experiments are never independent, so that any methods that assume such independence should be approached with great care, if at all. Three more realistic options have been most often considered in the literature: methods based on either permutations, on Simes' inequality, or assumption-free methods. Permutation-based methods have most to be said for them, as they adapt powerfully to the unknown joint distribution of the p-values and do not rely on the validity of the asymptotics of the raw p-values. Permutation methods are not completely assumption-free, however, and they are not available for all experimental designs. Computation time can also be a limiting factor when using them. Simes-based methods are an alternative, which can be assumed to be valid for the ubiquitous situation of two-sided test statistics that are at least asymptotically normal. Assumption-free methods can always be used, but generally sacrifice a lot of power relative to the other two types of methods, and should be considered only in situations that permutations and Simes-based methods are both excluded.

The type of multiple testing correction to be used depends crucially on the way the results are going to be used or reported. If individual rejected hypotheses are of interest, and if author or reader of the results is likely to take subsets of the collection of rejected hypotheses out of the context of the complete rejected set, then FWER controlling methods are advisable. If, on the other hand, the collection of rejected results as a whole is of interest, either because this entire collection is to be forwarded to a validation experiment, or because the overview is of interest, rather than the detailed results, then the FDP of this collection should be the focus, and FDR or FDP-type methods are preferable. FDP confidence methods are between FDR and FWER methods in many respects, sacrificing some of the power of FDR methods, but gaining more protection against the variability of FDP around FDR, and gaining the ability to make simultaneous statements about interesting subsets and supersets of the chosen results. These methods are especially useful if the hypotheses with top p-values are not necessarily all automatically of interest, but greater flexibility in decision-making is desired.

References

1. Dudoit S, Shaffer J, Boldrick J. Multiple hypothesis testing in microarray experiments. *Statistical Science* 2003; **18**(1):71–103.
2. Benjamini Y. Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal* 2010; **52**(6):708–721.
3. Farcomeni A. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research* 2008; **17**(4):347–388.
4. Roquain E. Type I error rate control for testing many hypotheses: a survey with proofs. *Journal de la Societé Française de Statistique* 2011; **153**(2):3–38.
5. Dudoit S, van der Laan M. *Multiple testing procedures with applications to genomics*. Springer Verlag, 2008.
6. Cox D. A remark on multiple comparison methods. *Technometrics* 1965; **7**(2):223–224.
7. Benjamini Y, Yekutieli D. False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association* 2005; **100**(469):71–81.
8. Ioannidis J. Why most published research findings are false. *PLoS medicine* 2005; **2**(8):e124.
9. Rothstein H, Sutton A, Borenstein M. *Publication Bias in Meta-Analysis*. Wiley Online Library, 2005.
10. Bender R, Lange S. Adjusting for multiple testing–when and how? *Journal of Clinical Epidemiology* 2001; **54**(4):343–349.
11. Goeman J, Solari A. Multiple testing for exploratory research. *Statistical Science* 2012; **26**(4):584–597.

12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995; **57**(1):289–300.
13. Storey J. The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* 2003; **31**(6):2013–2035.
14. Finner H, Roters M. On the false discovery rate and expected type I errors. *Biometrical Journal* 2001; **43**(8):985–1005.
15. Guo W, Bhaskara Rao M. On control of the false discovery rate under no assumption of dependency. *Journal of Statistical Planning and Inference* 2008; **138**(10):3176–3188.
16. Simes R. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**(3):751–754.
17. Sarkar S. Some probability inequalities for ordered mtp$_2$ random variables: a proof of the simes conjecture. *The Annals of Statistics* 1998; **26**(2):494–504.
18. Sarkar S. Two-stage stepup procedures controlling FDR. *Journal of Statistical Planning and Inference* 2008; **138**(4):1072–1084.
19. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* 2001; **29**(4):1165–1188.
20. Rødland E. Simes' procedure is 'valid on average'. *Biometrika* 2006; **93**(3):742–746.
21. Sarkar S. Fdr-controlling stepwise procedures and their false negatives rates. *Journal of Statistical Planning and Inference* 2004; **125**(1):119–137.
22. Reiner-Benaim A. Fdr control by the bh procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometrical Journal* 2007; **49**(1):107–126.
23. Kim K, Van De Wiel M. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics* 2008; **9**(1):114.
24. Yekutieli D. Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test* 2008; **17**(3):458–460.
25. Finner H, Dickhaus T, Roters M. On the false discovery rate and an asymptotically optimal rejection curve. *The Annals of Statistics* 2009; **37**(2):596–618.
26. Good P. *Permutation tests*. Wiley Online Library, 2000.
27. Pesarin F. *Multivariate permutation tests: with applications in biostatistics*. Wiley Chichester, 2001.
28. Huang Y, Xu H, Calian V, Hsu J. To permute or not to permute. *Bioinformatics* 2006; **22**(18):2244–2248.
29. Goeman J, Solari A. The sequential rejection principle of familywise error control. *The Annals of Statistics* 2010; **38**(6):3782–3810.
30. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273**(5281):1516–1517.
31. Hoggart C, Clark T, De Iorio M, Whittaker J, Balding D. Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology* 2008; **32**(2):179–185.
32. Van De Vijver M, He Y, Van't Veer L, Dai H, Hart A, Voskuil D, Schreiber G, Peterse J, Roberts C, Marton M, *et al.*. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine* 2002; **347**(25):1999–2009.
33. Rosenwald A, Wright G, Chan W, Connors J, Campo E, Fisher R, Gascoyne R, Muller-Hermelink H, Smeland E, Giltnane J, *et al.*. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* 2002; **346**(25):1937–1947.
34. Blanchard G, Dickhaus T, Hack N, Konietschke F, Rohmeyer K, Rosenblatt J, Scheer M, Werft W. μtoss—multiple hypothesis testing in an open software system. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, vol. 11, 2010.
35. Perneger T. What's wrong with Bonferroni adjustments. *British Medical Journal* 1998; **316**(7139):1236–1238.
36. Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 1967; **62**(318):626–633.
37. Rothman K. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; **1**(1):43–46.
38. Goeman J, Van De Geer S, De Kort F, Van Houwelingen H. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004; **20**(1):93–99.
39. Hummel M, Meister R, Mansmann U. Globalancova: exploration and assessment of gene group effects. *Bioinformatics* 2008; **24**(1):78–85.
40. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**(2):65–70.
41. Hochberg Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**(4):800–802.
42. Marcus R, Peritz E, Gabriel K. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**(3):655–660.
43. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988; **75**(2):383–386.
44. Hommel G. A comparison of two modified Bonferroni procedures. *Biometrika* 1989; **76**(3):624–625.
45. Westfall P, Young S. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, vol. 279. Wiley-Interscience, 1993.
46. Meinshausen N, Maathuis M, Bühlmann P. Asymptotic optimality of the Westfall–Young permutation procedure for multiple testing under dependence. *The Annals of Statistics* 2012; **39**(6):3369–3391.
47. Ge Y, Dudoit S, Speed T. Resampling-based multiple testing for microarray data analysis. *Test* 2003; **12**(1):1–77.
48. Phipson B, Smyth G. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology* 2010; **9**(1):39.
49. Benjamini Y, Hochberg Y. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* 2000; **25**(1):60–83.
50. Benjamini Y, Krieger A, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 2006; **93**(3):491–507.
51. Blanchard G, Roquain E. Adaptive false discovery rate control under independence and dependence. *The Journal of Machine Learning Research* 2009; **10**:2837–2871.
52. Sarkar S, Guo W, Finner H. On adaptive procedures controlling the familywise error rate. *Journal of Statistical Planning and Inference* 2012; **142**(1):65–78.
53. Blanchard G, Roquain E. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics* 2008; **2**:963–992.

54. Yekutieli D, Benjamini Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* 1999; **82**(1-2):171–196.
55. Romano J, Shaikh A, Wolf M. Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test* 2008; **17**(3):417–442.
56. Troendle J. Stepwise normal theory multiple test procedures controlling the false discovery rate. *Journal of Statistical Planning and Inference* 2000; **84**(1-2):139–158.
57. Ge Y, Sealfon S, Speed T. Some step-down procedures controlling the false discovery rate under dependence. *Statistica Sinica* 2008; **18**(3):881–904.
58. Korn E, Troendle J, McShane L, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* 2004; **124**(2):379–398.
59. Finner H, Dickhaus T, Roters M. Dependency and false discovery rate: asymptotics. *The Annals of Statistics* 2007; :1432–1455.
60. Troendle J. Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test* 2008; **17**(3):456–457.
61. Goeman J, Mansmann U. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* 2008; **24**(4):537–544.
62. Posch M, Zehetmayer S, Bauer P. Hunting for significance with the false discovery rate. *Journal of the American Statistical Association* 2009; **104**(486):832–840.
63. Storey J. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; **64**(3):479–498.
64. Storey J, Taylor J, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2004; **66**(1):187–205.
65. Schwartzman A, Lin X. The effect of correlation in false discovery rate estimation. *Biometrika* 2011; **98**(1):199–214.
66. Pounds S, Cheng C. Improving false discovery rate estimation. *Bioinformatics* 2004; **20**(11):1737–1745.
67. Owen A. Variance of the number of false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005; **67**(3):411–426.
68. Qiu X, Yakovlev A. Some comments on instability of false discovery rate estimation. *Journal of Bioinformatics and Computational Biology* 2006; **4**(5):1057–1068.
69. Efron B, Tibshirani R, Storey J, Tusher V. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001; **96**(456):1151–1160.
70. Qiu X, Klebanov L, Yakovlev A. Correlation between gene expression levels and limitations of the empirical bayes methodology in microarray data analysis. *Statistical Applications in Genetics and Molecular Biology* 2005; **4**(1):34.
71. Pawitan Y, Calza S, Ploner A. Estimation of false discovery proportion under general dependence. *Bioinformatics* 2006; **22**(24):3025–3031.
72. Efron B. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* 2007; **102**(477):93–103.
73. Efron B. Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association* 2010; **105**(491):1042–1055.
74. Goeman J, Solari A. Rejoinder. *Statistical Science* 2011; **26**(4):608–612.
75. Meinshausen N. False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics* 2006; **33**(2):227–237.
76. Lee W, Gusnanto A, Salim A, Magnusson P, Sim X, Tai E, Pawitan Y. Estimating the number of true discoveries in genome-wide association studies. *Statistics in Medicine* 2011; **31**(11–12):1177–1189.
77. Hommel G, Kropf S. Tests for differentiation in gene expression using a data-driven order or weights for hypotheses. *Biometrical journal* 2005; **47**(4):554–562.
78. Finos L, Salmaso L. Fdr-and fwe-controlling methods using data-driven weights. *Journal of Statistical Planning and Inference* 2007; **137**(12):3859–3870.
79. Hackstadt A, Hess A. Filtering for increased power for microarray data analysis. *BMC Bioinformatics* 2009; **10**(1):11.
80. van Iterson M, Boer J, Menezes R. Filtering, FDR and power. *BMC Bioinformatics* 2010; **11**(1):450.
81. Meinshausen N. Hierarchical testing of variable importance. *Biometrika* 2008; **95**(2):265–278.
82. Yekutieli D. Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association* 2008; **103**(481):309–316.
83. Benjamini Y, Bogomolov M. Adjusting for selection bias in testing multiple families of hypotheses. *Arxiv preprint arXiv:1106.3670* 2011; .
84. Goeman J, Finos L. The inheritance procedure: multiple testing of tree-structured hypotheses. *Statistical Applications in Genetics and Molecular Biology* 2012; **11**(1):1–18.
85. Zhong H, Prentice R. Correcting ‘winner’s curse’ in odds ratios from genomewide association findings for major complex human diseases. *Genetic Epidemiology* 2010; **34**(1):78–91.