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 on 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 previous generations 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.

Genomics data have proliferated and diversified. It is now possible to measure gene expression of messenger RNA or microRNA, DNA methylation, DNA polymorphisms and copy number variations, as well as related "omics" such as metabolomics. All these can typically be measured either with array-type technology or with next-generation sequencing techniques. In terms of multiple testing such data types are not extremely different. We focus on the multiple testing issues that are general to all genomics data types, although we will occasionally point out the additional problems and opportunities presented e.g. by discrete test statistics such as arise from sequencing. We will also be relatively brief on genome-wide association studies (GWAS), which have developed their own distinct multiple testing practice.

Earlier reviews on multiple testing that deal with genomics have appeared. We mention especially the excellent review by Dudoit, Shaffer and Boldrick [1], now a decade old, and the retrospective overview by Benjamini [2]. More

[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*

Prepared using simauth.cls *[Version: 2010/03/10 v3.00]*

technical overviews can be found in the papers of Farcomeni [3] and Roquain [4] and the book by Dudoit and Van der Laan [5]. This tutorial incorporates the developments of the last 10 years, taking on an applied perspective.

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. 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 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 hypothesis, as we'll see in Section 2.

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 bounded by 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. 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.

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 [8]. Selection of p-values occurs whenever many hypotheses are tested, and only a selected subset of those is reported or emphasized [9]. 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 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. 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 a very similar 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. It takes into account most of the selection that is done out of sight of other researchers, but ignores other selection mechanisms, such as *publication bias* [10]. In genomics, the natural family 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. We will assume that the researcher has only one hypothesis of interest for each probe.

1.2. Genomics, exploratory research 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 each of the probes on the microarray chip. In this experiment the number of hypotheses is in the order of 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 chosen not just on the basis of p-values or other statistical measures, but 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 the evidence for them must be provided by subsequent experiments [8]. This view, in which *anything goes* in exploratory research, turns out to be not 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 results 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 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, even though some biases inherent in the experimental design may be replicated in the validation experiment, especially if the same techniques are used in the 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. In situations in which full replication is not feasible, proper multiple testing in the exploratory phase may help give credibility to experiments that lack adequate validation.

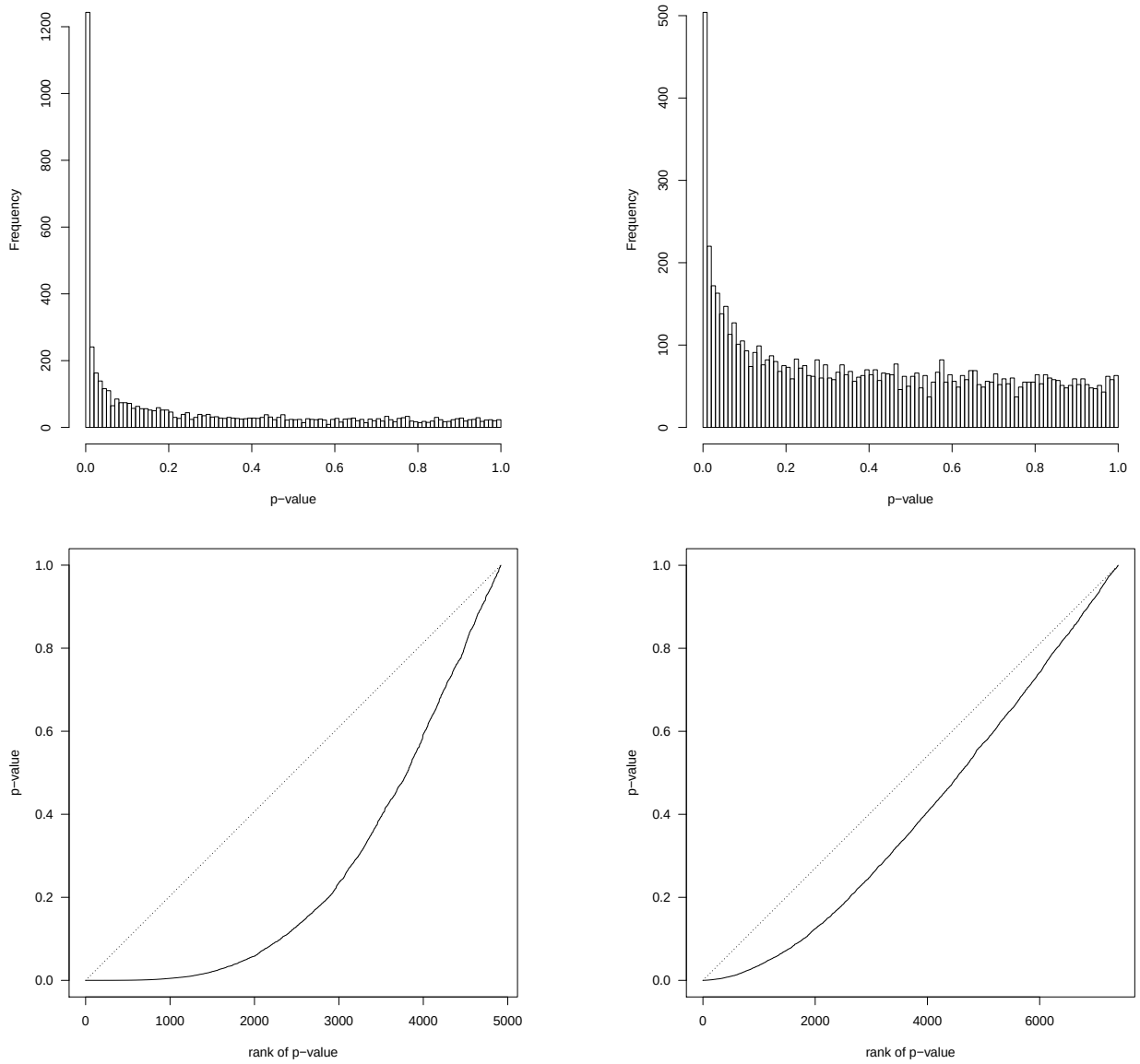
Multiple testing in genomics is also required in the validation experiments themselves if more than one finding is to be validated. Validation experiments are not exploratory but confirmatory experiments, and should be treated as such.

1.3. 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 [12]. This data set has gene expression profiles of 4,919 probes for 295 breast cancer patients. The second data set, of Rosenwald [13], 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 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 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.

In some types of studies, most notably GWAS, we expect a priori very few false null hypotheses. In such data a marked deviation from uniformity of the p-values such as we observe in Figure 1 is usually taken as an indication of lack of model fit or of inappropriately formulated null hypotheses, for example because important confounders, such as population structure, have been missed.



(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.

1.4. 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 algorithm of Box 1 on page 11. 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 [14]. For those who do not use R, SAS also has a large collection of multiple testing procedures [15]. Multiple testing in other commercial packages such as SPSS and Stata is, unfortunately, very limited.

An overview of the main R functions used to perform the most important multiple testing methods discussed in this paper is provided given in Table 2. We have included a commented R file with a examples of all the recommended methods that were described in this paper. For these examples we use a subset of the Van de Vijver data set described in Section 1.3.

2. Error rates

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 these terms, and by giving some guidance which type of method to use in which situation. Methods for FWER control are discussed in Section 4, and methods for FDR control in Section 5. FDP estimation and confidence is treated in Section 6.

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 null hypotheses, one for each genomic probe, which we would like to reject. An unknown number m_0 of these hypotheses are true, while the other $m_1 = m - m_0$ are false. We call the collection of true hypotheses $\mathcal{T} \subseteq \mathcal{H}$ and the remaining collection of false hypotheses $\mathcal{F} = \mathcal{H} \setminus \mathcal{T}$.The proportion of true hypotheses is denoted $\pi_0 = m_0/m$.

The goal of a multiple testing procedure is to choose a collection $\mathcal{R} \subseteq \mathcal{H}$ of hypotheses to reject. If we have p-values $p_1, \ldots, p_m$ for each of the hypotheses $H_1, \ldots, H_m$, an obvious choice is the collection $\mathcal{R} = \{H_i : p_i \leq T\}$, rejecting all hypotheses with a p-value below a threshold T. In this situation the multiple testing problem reduces to the choice of T. In some situations, however, rejected sets of other forms may be of interest (see Section 6.3).

Ideally, the set of rejected hypotheses $\mathcal{R}$ should coincide with the set $\mathcal{F}$ of false hypothese as much as possible. Two types of error can be made: false positives, or type I errors, are rejections of hypotheses that are not false, i.e. hypotheses in $\mathcal{R} \cap \mathcal{T}$; false negatives, or type II errors, are failed rejections of false hypotheses, i.e. hypotheses in $\mathcal{F} \setminus \mathcal{R}$. 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 multiple hypothesis testing procedure in a contingency table such as Table 1. We can observe the total number of hypotheses m and the number of rejected hypotheses $R = \#\mathcal{R}$, 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 a trade-off between the two must be made. If we reject more hypotheses we typically have more type I errors but fewer type II errors. Multiple testing methods try to reject as many hypotheses as possible while keeping some measure of type I errors in check. This measure is usually either the number V of type I errors or the *false discovery proportion* (FDP) Q, defined as

$$Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{otherwise,} \end{cases}$$

which is the proportion of false rejections among the rejections, defined as 0 if no rejections are made.

Most multiple testing methods choose the threshold T as a function of the data so that the set $\mathcal{R}$ of rejected hypotheses is random, and so both V and Q are random variables. In this situation we cannot keep the values of V and Q themselves small, but must focus on relevant aspects of their distribution. Different types of multiple testing methods focus on different summaries of the distribution of V and Q. The most popular methods either control the *familywise error* (FWER), given by

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

or the *false discovery rate* (FDR) [16], given by

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

FWER focuses on the probability that the rejected set contains any error, while FDR looks at the expected proportion of errors among the rejections. FWER or FDR is controlled at level α, which 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.

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 method is automatically also an FDR-controlling method. 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 methods to have more power than FWER-based ones. In practice, FDR controlling methods are especially more powerful than FWER controlling methods if there are many false hypotheses. Conversely, if all hypotheses are true, FDR and FWER are identical; since $R = V$ in this case, Q is a Bernoulli variable, and $\mathrm{E}(Q) = \mathrm{P}(Q > 0)$. 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 $H \in \mathcal{R}$. This difference has important practical consequences. 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 $\mathcal{S} \subset \mathcal{R}$. This means that if a FWER-controlling procedure has rejected a set $\mathcal{R}$ that is 'too large' e.g. for the ensuing validation experiment, a user may take any subset $\mathcal{S}$ of the discoveries and still control FWER. The corresponding property does not hold for FDR control. In fact, it was shown by Finner and Roters [17] 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 holds for the full set $\mathcal{R}$ only, and does not translate to subsets $\mathcal{S}$ or to individual hypotheses in $\mathcal{R}$. This subsetting property, or the lack of it, has implications for the way FWER and FDR can be used, and we come back to this in Sections 4.5 and 5.4.

To compare FDR, FWER and unadjusted testing it is helpful to look a bit more closely at the rows and columns of the contingency Table 1. In the contingency table, we can look at the number of type I errors V standardized by column, by row, or unstandardized. Standardized by column, we look at the number of type I errors per true hypothesis V/m_0. The expectation of this quantity is kept below α if we do unadjusted testing. Standardized by row, we look at the number of type I errors per rejected hypothesis V/R. We keep the expectation of this quantity below α by controlling FDR. Row and column ratios in contingency tables are related to each other through Bayes' rule, and the relationship depends crucially on the prevalence $\pi_0 = m_0/m$ of true hypotheses. At very high prevalence π_0, the row ratio V/R is larger than the column ratio V/m_0, and consequently control of FDR must be more stringent than unadjusted testing. Conversely, if the prevalence π_0 is very low, the column ratio V/m_0 is larger than the row ratio V/R, and FDR control may be less stringent than unadjusted testing. We conclude that the stringency of the FDR criterion depends very much on the proportion of hypotheses that is true. At realistic high values of π_0 FDR is much more stringent than unadjusted testing, but FDR may also, at least in theory, be less stringent than unadjusted testing. It is a reasonable additional demand of methods that control FDR that they also keep $\mathrm{E}(V/m_0) \leq \alpha$, by never rejecting hypotheses with p-values above α, and many of them conform to this. In contrast to FDR control and unadjusted testing, FWER control keeps V itself small, without standardizing by row or column.

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 [18]. We come back to this view of FDR in Sections 6.1 and 6.2. Empirical Bayesian FDR in turn links closely to fully Bayesian approaches to multiple testing [19].

In methods that control an error rate, such as FDR or FWER, the user chooses an error rate to be controlled, and the multiple testing method finds a rejected set $\mathcal{R}$ for the user according to the criterion. This contrasts with FDP confidence methods that let the user choose the set $\mathcal{R}$ freely, and which subsequently try make a confidence interval the FDP of that set. In this type of methods the set $\mathcal{R}$ is not a random quantity, and the number of errors V or false discovery proportion Q in this set are fixed but unknown quantities, which can in principle be estimated. In practice, of course, the set $\mathcal{R}$ to be rejected is not determined before data collection, but will often be chosen in some partially data-dependent way. Any estimates and confidence statements for Q or V need to be corrected for bias resulting from such a data-dependent choice. We will look into such FDP estimation and FDP confidence methods in greater detail in Section 6, and explain the difference with FDR-control.

3. Assumptions of multiple testing methods

In statistics, stronger assumptions generally allow more powerful statements. In multiple testing, the most crucial assumptions to be made concern the dependence structure of the p-values. Much work has been done under the assumption of independent 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 making a certain assumption on the dependence structure of the p-values, known as the PDS assumption. The third kind uses permutations to adapt to the dependence structure of the p-values. All three types of assumptions have been used in methods for FWER control, FDR control and FDP estimation. Table 2 gives an overview of the methods that use each of the different assumptions to achieve control of each of the error rates. Before we move on to specific methods, we first discuss the assumptions in some detail. This will be the most technical part of this tutorial.

Assumptions	Error criterion		
	FWER control	FDR control	FDP confidence
No assumptions	Holm `p.adjust` Sections 4.1, 4.2	Benjamini & Yekutieli `p.adjust` Section 5.2	Goeman & Solari `pickSimes` (*cherry*) Section 6.3
PDS assumption	Hommel `p.adjust` Section 4.3	Benjamini & Hochberg `p.adjust` Section 5.1	Goeman & Solari `pickSimes` (*cherry*) Section 6.3
Permutations	Westfall & Young `mt.maxT` (*multtest*) Section 4.4	 Section 5.3	Meinshausen `howmany_dependent` (*howmany*) Section 6.4

Table 2. Overview of error criteria and assumptions considered in thus tutorial, with the main representative method for each combination, if present. The table mentions the name of the method and an `R` function (with required package) that implements it, and refers to the appropriate section of this tutorial. The methods of Storey (Section 6.1) and Efron (Section 6.2) fall outside this classification scheme.

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. Assumptions on the p-values often involve only the p-values of true hypotheses. We denote these by $q_1, \ldots, q_{m_0}$. By the definition of a p-value, if their corresponding hypotheses are true, these p-values are either uniformly distributed between 0 and 1, or they can be stochastically greater than uniform if data are discrete: we have, for $i = 1, \ldots, m_0$,

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

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 says that simultaneously, with probability at least $1 - \alpha$,

$$q_i > \alpha/m_0 \qquad \text{for all } i = 1, \ldots, m_0.$$

The Bonferroni inequality is discussed in detail in Section 4.1. The second inequality is due to Hommel [20], 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 the true hypotheses. Hommel's and Bonferroni's inequalities are valid whatever the dependence of p-values, as long as (1) holds.

Probability inequalities have a 'worst case' distribution for which the inequality is an equality, but they 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 [21], which can be quite conservative for more realistic distributions.

To avoid having to cater for exotic worst case distributions, assumptions can be made to exclude them. In particular, an assumption of *positive dependence through stochastic ordering* (PDS) can be made, which excludes the 'worst case' distributions of the inequalities of Hommel and Bonferroni. Technically, the PDS condition comes in two forms [22, 23]. The weaker PDS condition says that

$$\mathrm{E}[f(q_1, \ldots, q_{m_0}) \mid q_i = u] \tag{3}$$

is non-decreasing in u for every i and for every coordinate-wise non-decreasing function f. The stronger PDS condition is identical except that (3) is replaced by

$$\mathrm{E}[f(p_1, \ldots, p_m) \mid q_i = u], \tag{4}$$

so that it involves not only non-decreasing functions of p-values of true hypotheses but also of false ones. The second PDS condition is slightly more restrictive, but for all practical purposes we can view the two PDS conditions as identical. In Table 2, the FWER controlling methods and FDP confidence methods require the weaker PDS condition, while some FDR controlling methods, most notably the famous Benjamini & Hochberg procedure (Section 5.1), require the stronger one.

The weaker PDS condition is a sufficient condition for a probability inequality due to Simes [24]. This Simes inequality is related to Hommel's inequality, but more powerful. It says that with probability at least $1 - \alpha$, simultaneously

$$q_{(i)} > \frac{i\alpha}{m_0} \qquad \text{for all } i = 1, \ldots, m_0. \tag{5}$$

Simes' inequality improves upon both Hommel's and Bonferroni's inequalities. The critical values of Simes' inequality are larger than those of Hommel's inequality by a factor $\sum_{j=1}^{m_0} 1/j$, which, for large m_0, is approximately equal to $\log(m_0) + \gamma$, where $\gamma \approx 0.577$ is the Euler-Mascheroni constant. As a probability inequality, 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 [24], which is relatively unexotic.

Examples of cases under which the PDS condition holds include one-sided test statistics that are jointly normally distributed, if all correlations between test statistics are positive, or two-sided joint normal test statistics under a different condition that allows some negative correlations; Simes' inequality has also been shown to hold for some important cases for which PDS fails, such as the for certain basic multivariate t-distributions [25]. Even though these conditions are not guaranteed to hold for all distributions relevant for genomics, methods based on the PDS condition turn out to be quite robust in practice. This has been corroborated theoretically by Rødland [26], who wrote that "distributions for which Simes' procedure fails more dramatically must be somewhat bizarre". The same robustness has been demonstrated in many simulation experiments [24, 27, 28, 29]. 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 use methods based on the PDS condition [30].

A third 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 [31] and Pesarin [32]. Permutation tests have two great advantages, both of which translate to permutation-based multiple testing. First, they give exact error control without having to rely on asymptotics for the assumption (1), allowing reliable testing even when asymptotic p-values are unreliable. Second, permutation tests do not use any probability inequality, but generally 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. When test statistics are discrete, such as in sequencing or GWAS data, permutation-based multiple testing can also automatically eliminate hypotheses that cannot possibly obtain small p-values, resulting in higher power for the other probes [33].

Unfortunately, valid permutation tests are not defined for all hypotheses in all models; permutation tests are not assumption-free [34]. 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 [35]. 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.

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 is often assumed that local correlations between adjacent hypotheses along the genome are so strong that testing the whole genome is equivalent to performing only 10^6 tests, regardless of the number of hypotheses actually tested. Under this assumption, FWER can be controlled using a fixed genome-wide significance level of $\alpha = 5 \times 10^{-8}$, corresponding to a Bonferroni adjustment (see Section 4.1) for 10^6 tests. 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 [36].

4. Methods for FWER control

The most classical way of controlling for multiple testing is by FWER control, and consequently this is the type of multiple testing for which most methods are available. We follow the outline sketched in the previous section. We first discuss the methods of Bonferroni and Holm, which are valid under any dependence of the p-values, then Hochberg's and Hommel's methods, which are valid if the PDS assumption holds, before looking at the permutation-based method of Westfall & Young.

4.1. Bonferroni

The method of Bonferroni controls FWER at level α by rejecting hypotheses only if they have raw p-value smaller than α/m. This single-step adjustment of the significance threshold is the simplest, oldest, and most well-known multiple testing method, and is attractive because of its simplicity. However, it is also known to be conservative, especially if many hypotheses are false, or if strong positive correlations between p-values occur.

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. One of the most widespread of these misunderstandings is that the method would be based on an assumption of independence between p-values [37]. This misunderstanding comes from a frequently used, but deficient motivation for Bonferroni, saying that the probability of making a false rejection if all m_0 p-values of true 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 $m_0\alpha/m \leq \alpha$. By this reasoning, Bonferroni seems like a method that only provides approximate FWER control, and that requires an assumption of independence for its validity. In fact, the method of Bonferroni provides exact FWER control under any dependence structure of the p-values.

To properly motivate the Bonferroni method we should look at it as 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 hypotheses, that the probability that there is some i for which $q_i \leq \alpha/m$ is given by

$$\mathrm{P}\big(\min_i q_i \leq \alpha/m\big) = \mathrm{P}\big(\bigcup_{i=1}^{m_0}\{q_i \leq \alpha/m\}\big) \leq \sum_{i=1}^{m_0} \mathrm{P}(q_i \leq \alpha/m) \leq m_0\frac{\alpha}{m} \leq \alpha. \tag{6}$$

Since the method of Bonferroni only commits a type I error if $q_i \leq \alpha/m$ for some i, this proves FWER control at level α for the Bonferroni method.

A few things can be learnt from this derivation. In the first place, the FWER control of Bonferroni is not approximate, and it is valid for all dependence structures of the underlying p-values. Secondly, the three inequalities in (6) 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 hypotheses, Bonferroni will be conservative. The middle inequality, that uses (1), says that Bonferroni is conservative if the raw p-values are, which happens if test statistics are discrete [38]. 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, i.e. if the p-values have strong negative correlations.

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 critical values $1-(1-\alpha)^{1/m}$ that can be calculated for independent p-values [39]. For $m = 5$ and $\alpha = 0.05$ we find a critical value of 0.01021 under independence 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 $m_0 = m$ [37, 40]. This type of control is known as *weak* FWER control. On the contrary, as we can see from (6) 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, because the probability of a type I error occurring is at most α if all hypotheses are true, but only in that situation. Related to weak control methods but less overconfident are global testing methods [41, 42] 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.

4.2. Holm

Holm's method [43] 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 without any assumptions aside from (1). Holm's method should, therefore, always be preferred to Bonferroni, except possibly for reasons of simplicity.

Holm remedies part of the conservativeness in the Bonferroni method arising from the right-hand inequality of (6), which makes Bonferroni control FWER at level $\pi_0\alpha$ rather than α. 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 the 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 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 Box 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 $p_{(1)}, \ldots, p_{(m)}$.
2. Multiply each $p_{(i)}$ by its adjustment factor a_i, $i = 1, \ldots, m$, 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:
   $$\tilde{p}_{(i)} = \max_{j=1,\ldots,i} a_j p_{(j)}$$
   (b) *Step-up (all others)*: Decrease the highest p-value in all violating pairs:
   $$\tilde{p}_{(i)} = \min_{j=i,\ldots,m} a_j p_{(j)}$$
4. Set $\tilde{p}_{(i)} = \min(\tilde{p}_{(i)}, 1)$ for all i.

Box 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.

4.3. 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 to exclude this worst case a priori, and as a result gain some power. One such assumption could be that the PDS condition holds for the subset of true hypotheses (Section 3). This assumption makes the use of Simes inequality (5) inequality possible, and therefore the use of Hochberg's and Hommel's methods. Hommel's method is always the more powerful of the two, but Hochberg's method is simpler.

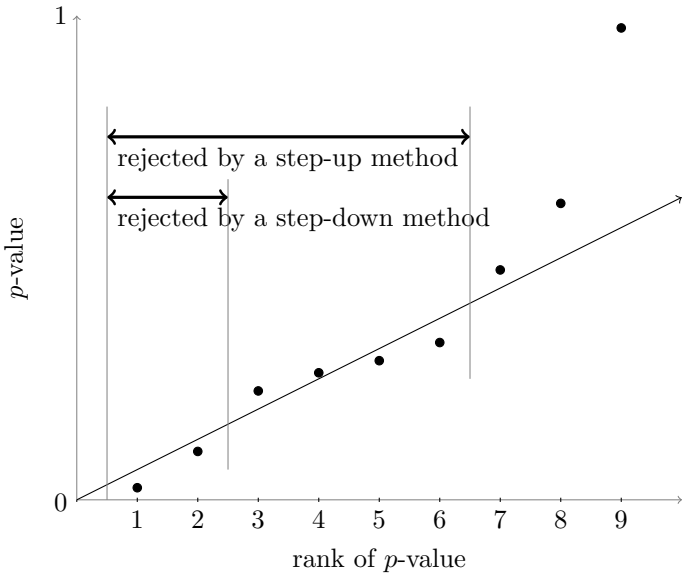
Hochberg's method [44] is very similar to Holm's method, but more powerful. Hochberg's method (not to be confused with Benjamini & Hochberg's method, Section 5) compares each ordered p-value $p_{(i)}$ to a critical value $\alpha/(m-i+1)$, the same as Holm's. It then finds the largest j such that $p_{(j)}$ 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 $p_{(1)}, \ldots, p_{(m)}$ and the critical value curve $\alpha/m, \alpha/(m-1), \ldots, \alpha$ as its critical value, while Hochberg's uses the last 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.

Hommel's procedure [45] is an alternative FWER-controlling procedure that is valid under the same PDS assumption. It always rejects at least as much as Hochberg's, and possibly more [46], but is more complicated and computationally more demanding. However, 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. In practice, the difference between Hommels'and Hochberg's methods is usually small. It is worth mentioning that Hommel's method is unrelated to Hommel's probability inequality (Section 3). Proof of FWER control of Hochberg's and Hommel's methods is based on a combination of the Simes inequality with the powerful *closed testing procedure* [47], which is an important instrument of method development in multiple testing.

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.



4.4. 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 this way, the conservativeness associated with probability inequalities such as Boole's or Simes' can be avoided, and instead the multiple testing procedure can adapt to the true null distribution of the p-values. The permutation-based FWER-controlling procedures of Westfall & Young [48] are often more powerful than the methods of the previous sections, and have even been shown to be asymptotically optimal for a broad class of correlation structures [49].

In the methods of Westfall & Young, permutations are used to find the α-quantile of the distribution of the minimum p-value of the m_0 true hypotheses. This same quantile is bounded from below by α/m by Bonferroni,

but, as we have seen in Section 4.1, this bound is often conservative. The methods of Westfall & Young use permutations to obtain a more accurate threshold, which is usually larger than α/m.

As explained in Section 3, we need valid permutations to do permutation testing. 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 consider carefully whether 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.

Two variants of Westfall & Young's methods exist: the max$T$ and min$P$ methods. The max$T$ method uses the raw p-values directly as input for the method, while the min$P$ method transforms these p-values to (raw) permutation p-values first.

Practically, the max$T$ method of Westfall & Young starts by making k permuted data sets, and recalculating all m raw 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 $\mathbf{P}$ 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 $\mathbf{P}$ 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 their max$T$ method, except that instead of the raw p-values it uses the per-hypothesis permutation p-values in the matrix $\mathbf{P}$. A fast algorithm for this procedure was designed by Ge, Dudoit and Speed [50]. Since permutation p-values take only a limited number of values, the matrix $\mathbf{P}$ will always contain many tied values, which is an important practical impediment for the min$P$ method, as we'll see below.

The number of permutations is always an issue with permutation-based multiple testing. In data with a small sample size this number is necessarily limited, but it quickly becomes very large 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 used. Additional randomness is introduced in this way, which makes rejections and adjusted p-values random, especially with only few random permutations. 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, although some computational tricks are possible here [51].

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 $\mathbf{P}$ 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 expected 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 wonderful, it is still sensible to be careful with this, 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 can be 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, guaranteeing a uniform distribution of the input p-values for the method.

The max$T$ method has a great advantage over other multiple testing procedures in the situation that the tests for the different hypotheses are very diverse in terms of power. This occurs especially in discrete data, such as e.g. in sequencing experiments. In such situations the max$T$ method will automatically put more focus on the hypotheses with more power, thus increasing the overall power of the procedure. To see how this works, consider a simple example of a multiple testing procedure in which m binomial tests have been performed, but each with different sample size. The tests with smaller sample size will not achieve a very small p-values in the real data, but will also never achieve very small p-values in the permuted data. When calculating the minimum p-value distribution, these tests therefore hardly play a role: they are effectively ignored by the first steps of the max$T$ procedure. In effect, the max$T$ procedure severely downweights tests with low power, gains higher weight for highly powered tests in return, and achieves higher power for the complete procedure as a result [38]. This property is specific for the max$T$ procedure, and is not shared by the min$P$ variant.

For genomics data analysis permutation-based multiple testing is recommended whenever suitable permutations are available, and the sample size is large enough to allow permutation testing. In such situations, these methods combine guaranteed FWER control with typically higher power than Holm's or Hommel's methods, especially if p-values are strongly dependent. The min$P$ method is computationally prohibitive in larger problems because of the large number of permutations needed, but the max$T$ method is almost always feasible.

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 [52]. 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.

Software for the Westfall & Young procedures is somewhat limited. For some hypothesis tests the max$T$ and min$P$ procedures are implemented in the `mt.maxT` and `mt.minP` procedures in the `multtest` package in `R`. If the user is prepared to make the permutation matrix $\mathbf{P}$ manually, the `flip` package has a generic max$T$ and min$P$ implementation in the `flip.adjust` function. More is available in SAS PROC MULTTEST.

4.5. Use of FWER control

Since the advent of genomics data, FWER 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 suggest 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 considerable, 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 end-stage experiments 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. FWER control, therefore, is especially relevant for the validation experiments themselves. Whenever possible, powerful methods such as Westfall & Young's should be used to negate any conservativeness of FWER-based methods.

5. Methods for FDR control

The seminal paper in which Benjamini & Hochberg [16] introduced the concept of FDR has changed thinking about multiple testing quite radically, showing that FWER control is not the only way to do 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. Like for FWER-based methods, we discuss methods that are generally valid, methods valid under the PDS assumption, and methods based on permutations. However, for FDR, unlike for FWER, the PDS-based method is the oldest and best known one, so we start there.

5.1. Benjamini & Hochberg

The Benjamini & Hochberg procedure [16] 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

$$c_i = i\alpha/m,$$

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 (5), and the critical values of the Benjamini & Hochberg procedure are those of Simes' inequality with $m = m_0$.

It has been shown [53, 22] that this procedure controls FDR at level α under the stronger PDS assumption (4). This is virtually 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 3, the validity of the procedure seems quite robust, at least for two-sided tests that are asymptotically normal. In fact, control of FDR under the PDS assumptions is even at level $\pi_0\alpha$, rather than at level α, where $\pi_0 = m_0/m$, the proportion of true hypotheses among the hypotheses.

The critical values of the Benjamini & Hochberg procedure are much larger than those of Hochberg or Hommel, so that many more rejections can usually 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 1.12×10^{-5} for Hommel'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 Hommel. 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 Hommel'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 denominator 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 the proportion of true hypotheses 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 [54]. 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 [55], 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 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 [55, 29, 56]. 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 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 2 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 [57]. 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. Ties in these adjusted p-values are very frequent. 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 6.1). Some care must be applied when interpreting adjusted p-values based on FDR control, however, as we'll discuss in Section 5.4.

5.2. FDR control under general dependence

If the PDS assumption underlying the Benjamini & Hochberg procedure is not valid, an alternative is the procedure of Benjamini & Yekutieli [53]. 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 (5) itself. It is a step-up procedure that compares each ordered p-value $p_{(i)}$ with the critical value

$$c_i = \frac{i\alpha}{m\sum_{j=1}^{m} 1/j},$$

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. The Benjamini & Yekutieli procedure is more conservative than the Benjamini & Hochberg procedure because the critical values are decreased 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 many hypotheses 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 (Section 4.2), 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 [23] and by Blanchard and Roquain [58]. The latter authors proved that any step-up method with critical values of the form

$$c_i = \frac{\alpha}{m}\sum_{j=1}^{i} j f_j, \tag{7}$$

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 structure 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 [23], 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 [58]. 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.

5.3. 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 4.4) has not yet been found.

Research in this area is ongoing. The subject was pioneered by Yekutieli and Benjamini [59] who suggested a permutation-based procedure but without full proof of FDR control. Romano, Shaikh and Wolf [60], building on earlier work by Troendle [61] that had restrictive parametric assumptions, proposed a method that uses the bootstrap instead of permutations to control FDR asymptotically. Ge, Sealfon and Speed [62] 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 than Westfall & Young, and 4 in the Rosenwald data, like Westfall & Young. None of the permutation FDR methods comes with user-friendly software.

A different option is to do the procedure of Benjamini & Hochberg on permutation p-values. Such permutation testing does not adapt to the unknown correlation structure of the p-values, but is valid if the PDS assumption holds for the permutation p-values. Like the combination of permutation testing with Holm's or Hommel's methods in Section 4.4, power of such a procedure may suffer unless a large multiple of m/α permutations are used, although the step-up nature of the Benjamini & Hochberg procedure alleviates this problem if many hypotheses are false. Some computational tricks are available to speed up calculation [51].

5.4. 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 [63, 22, 64, 65], sometimes heavily [17]. 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 the FDP Q, which is a variable quantity because the rejected set $\mathcal{R}$ is random. It has been pointed out, however, that the actual value of Q realized by FDR-controlling procedures can be quite variable, especially when p-values are dependent. 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 0.10 can, for example, be greater than 0.29 more than 10% of the time even under independence [63], and dependence between p-values can increase this variability dramatically [22]. 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 2, 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 this does not translate to subsets of $\mathcal{R}$ or to specific individual hypotheses that are elements of $\mathcal{R}$ [17]. 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 [17]. 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 from the complete set of results, 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 [65]. For correct interpretation of FDR control results, and to prevent inadvertent cheating, it is important to keep the rejected set intact.

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 reasonably 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 that 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 [66]. 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.

6. Methods for FDP estimation and confidence

Next to methods for FDR control there are also methods that approach FDP from an estimation perspective, or which construct confidence bounds for FDP. Since both FDR control and FDP estimation and confidence aim to

say something about the false discovery proportion Q, we will discuss the difference between FDR control and FDP estimation and confidence briefly before turning to specific methods.

To understand the difference between FDR control and FDP estimation, imagine a researcher repeating the same experiment many times. If the researcher used an FDR controlling method to analyze the experiment, he or she would find a new set of rejected hypotheses every time, obtaining sets $\mathcal{R}_1, \ldots, \mathcal{R}_z$ in z experiments. Each of these sets would have its own unknown false discovery proportion, say $Q_1, \ldots, Q_z$. Control of FDR means that, if z is large, the average of the Q_i's is bounded by α. Since the rejected sets are different upon every repetition of the experiment, FDR control, clearly, is a property of the procedure leading to a rejected set $\mathcal{R}_i$, not of the rejected set $\mathcal{R}_i$ itself [64]. Of course, the fact that $\mathcal{R}_i$ was chosen by a method with FDR control says something about the properties of the set $\mathcal{R}_i$, but only indirectly.

FDP estimation and confidence methods look at the same problem in a different way. A researcher using such a method investigates a set $\mathcal{R}$ and obtains an estimate $\hat{Q}$, or a confidence interval $[L, U]$, for the false discovery proportion Q of that set. Upon repeating the experiment, it is assumed that the researcher investigates the same set $\mathcal{R}$ every time, but obtains new estimates $\hat{Q}_1, \ldots, \hat{Q}_z$ and confidence intervals $[L_1, U_1], \ldots, [L_z, U_z]$. The estimates $\hat{Q}$ are estimates in the classical sense of the fixed parameter Q belonging to the fixed set $\mathcal{R}$, and can have classical properties such as consistency and unbiasedness. The confidence intervals should have the coverage property that, if z is large, the true value of Q is contained in the confidence interval $[L_i, U_i]$ at least a proportion $1 - \alpha$ of the time. Since these FDP estimates and confidence statements are about the same fixed set $\mathcal{R}$, they are directly informative about $\mathcal{R}$ itself.

An important practical difference between FDR control and FDP estimation and confidence is that the routine of the multiple testing procedure is reversed. FDR control starts with the choice of the error rate to be controlled, and the procedure that controls it, and finds a single rejected set that fulfils the criteria. In contrast, FDP estimation and confidence starts with the set of hypotheses the researcher would like to reject, and finds an estimate or confidence interval for the error rate (FDP) of that set.

The potential for FDP estimation, as opposed to FDR or FWER control, is therefore that it allows researchers to approach multiple testing in a different way that reverses the traditional roles of the user and the multiple testing procedure [11]. If reliable estimates of Q are available for every set $\mathcal{R}$ of potential interest, the user can review these sets and their estimates, and select the most promising one. The role of the multiple testing procedure can be to inform the user of the likely FDP in the set by providing an estimate and a confidence interval for Q. Alternatively, methods may find estimates and confidence intervals for the number of type I errors V, rather than for the false discovery proportion Q.

Allowing the user to choose the rejected set $\mathcal{R}$ also opens the possibility of choosing sets that do not necessarily consist of the hypotheses with top p-values. This can be valuable in exploratory genomics research, in which a mixture of statistical and biological considerations may play a role in deciding which hypotheses are candidates for rejection. In such situations it can be important for a researcher to obtain a reliable estimate or confidence statement about the actual set $\mathcal{R}$ that is chosen.

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 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 used for FDP estimation, and this set will probably be selected because it has a small estimate $\hat{Q}$. The value of $\hat{Q}$ for the post hoc selected $\mathcal{R}$ is therefore very likely to be underestimated. Honest estimation of Q should protect against this estimation bias or correct for it. The second difficulty is to provide an assessment of the uncertainty in an estimate $\hat{Q}$. 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 $\mathcal{R}$. Proper confidence statements should take this variability into account.

We focus in this section mostly on methods that do not only provide a point estimate of Q, but also assess its variability by providing confidence statements for FDP. As with FWER and FDR control we discuss PDS-based and worst-case methods (Section 6.3) as well as permutation methods (Section 6.4). Before going into such methods, we briefly review two estimation approaches that take an empirical Bayes perspective: Storey's (Section 6.1) and Efron's (Section 6.2). Both methods provide point estimates only.

6.1. Storey's approach and SAM

Simple and intuitive FDP point estimates for the collection of the top k hypotheses with best p-values can be based on the p-value rank plots and histograms of Figure 1. This was first suggested by Storey [67], who was motivated by the empirical Bayesian view of FDR [18]. Storey considered only sets of rejected hypotheses of the

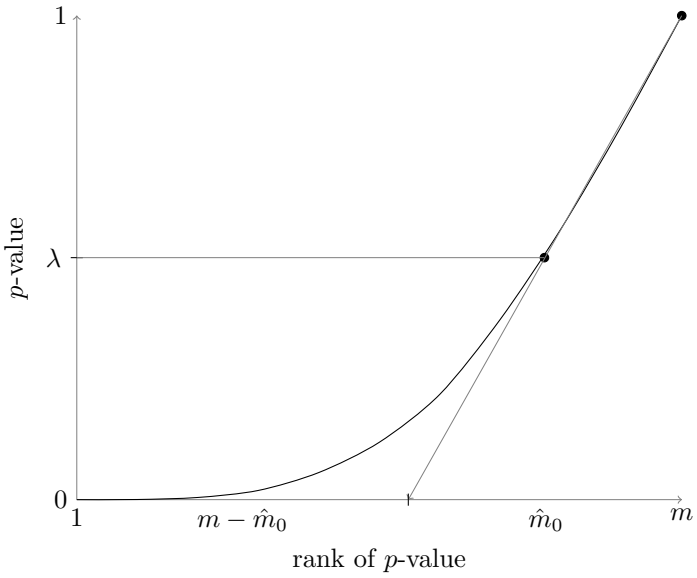
form $\mathcal{R} = \{H_i : p_i \leq t\}$ for some constant t. In such a rejected set, the expected number of true hypotheses to be rejected is $m_0 t$, because p-values of true hypothesis follow a uniform distribution. This number can be estimated by substituting an estimate $\hat{m}_0$ for m_0. Storey suggests

$$\hat{m}_0 = \frac{\#\{p_i > \lambda\} + 1}{1 - \lambda}, \tag{8}$$

where $0 \leq \lambda < 1$ is some constant. The value of λ is typically taken as $1/2$, although $\lambda = \alpha$ has also been advocated [56]. 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 estimate (**??**) is illustrated graphically in Figure 3. 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

$$\hat{Q} = \frac{\hat{m}_0 t}{\#\mathcal{R}} = \frac{t(\#\{p_i > \lambda\} + 1)}{(1 - \lambda)\#\{p_i \leq t\}}. \tag{9}$$

Figure 3. Illustration of Storey's estimate of m_0. The diagonal line connects the point $(m, 1)$ and the point at which the curve is equal to λ. The point at which it splits the x-axis determines the estimated number of true hypotheses.



In the Van de Vijver data, Storey estimates $\hat{m}_0 = 2230$ ($\hat{\pi}_0 = 0.45$). For the 1340 hypotheses rejected by the Benjamini & Hochberg procedure, this leads to an estimated FDP of only 2.3%. In the Rosenwald data, the same method gets $\hat{m}_0 = 5632$ ($\hat{\pi}_0 = 0.76$), which results in an estimated FDP of 3.8% for the 72 hypotheses rejected by the Benjamini & Hochberg procedure. Through its incorporation of a π_0 estimate, Storey's method is especially more optimistic than the Benjamini & Hochberg procedure in data sets with strong effects.

Storey's estimate has been derived under the assumption of independence among p-values. At first sight, however, it appears hardly affected by dependence among p-values. The expected number of true 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 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 [68], which says that $\lim_{m \to \infty} \hat{Q} \geq Q$. 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 [69]. 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 [70, 71, 72, 29, 60, 56]. 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 [71, 69], 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 (9) is below α. This actually links quite closely to the Benjamini & Hochberg procedure. If instead of using (8) we estimate m_0 conservatively by m, then we have $\hat{Q} = mt/(\#\{p_i \leq t\})$. 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 5.1. There is no proof of FDR control for Storey's procedure under realistic dependence, and simulations indicate lack of FDR control in such situations [70, 72, 29, 60, 56]. In the Van de Vijver data control of FDR using Storey's method would lead to 1787 rejections with $\lambda = 1/2$, or 1564 with the less anti-conservative $\lambda = \alpha$ as advocated by Blanchard and Roquain [56]. In the Rosenwald data, the same procedures lead to 133 and 100 rejections.

The method known as Significance Analysis of Microarrays (SAM) [73] can be viewed as a permutation based variant of Storey's method. It has been shown, both using theory [1], and by simulations [29] that SAM does not control or reliably estimate FDP or FDR, but instead estimates the less meaningful quantity $\mathrm{E}(V)/\mathrm{E}(R)$. Since this quantity is easier to control, SAM tends to be much less stringent than FDR-controlling methods.

6.2. Efron's approach

Storey's approach to FDP estimation can only make statements about the FDP of a set containing the top k p-values. As explained in Section 5.4, such statements do not translate to individual hypotheses within such a set. Storey's approach is also not usable for sets of different form. The *local FDR* approach of Efron [74] remedies this by providing point estimates for FDP of arbitrary sets and for individual hypotheses. This remedy comes at the cost of additional assumptions.

Where Storey was already motivated by the empirical Bayes approach to FDR, Efron [74] 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 $S_1, \ldots, S_m$ corresponding to hypotheses $H_1, \ldots, H_m$ were randomly drawn from a mixture distribution with density

$$f(t) = \pi_0 f_0(t) + (1 - \pi_0) f_1(t). \tag{10}$$

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 (10) 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

$$\mathrm{P}(H_i \text{ is true} \mid S_i = t) = \frac{\pi_0 f_0(t)}{f(t)}, \tag{11}$$

which Efron calls the *local FDR*. This is a powerful FDR-type statement about an individual hypothesis. By estimating π_0 and $f(t)$ from the data, local FDR may be determined for every hypothesis.

At first sight, the assumption (10) hardly seems an assumption at all. In fact, plotting a histogram of the observed test statistics coming from a genomics experiment will often produce a distribution that looks very much like the one suggested by (10). On closer inspection, however, it becomes apparent that this assumption is not without cost. We can say that truth or falsehood of hypotheses in Efron's model becomes a random, rather than a fixed effect. Whether such a random effects view of hypotheses is sensible can be disputed [75, 76]. Typically, random effects are used when we are not interested in individual entities but only in the population they are drawn from. This does not apply to hypotheses in genomics, which generally represent genes or similar entities that are individually meaningful, and are reported as such.

The exchangeability assumption on the hypotheses made in (10) relates to the 'cheating with FDR' discussion we touched upon in Section 5.4, and which arose when hypotheses are not homogeneous or exchangeable. It helps to compare the local FDR with an FDR adjusted p-value, which seems its closest analogue so far. On the one hand, because local FDR estimates condition on $S_i = t$ rather than $S_i \geq t$, the evidence of hypotheses with larger

test statistics is not directly mixed with the evidence for the hypothesis itself, as happens with FDR adjusted p-values. On the other hand, the cheating with FDR phenomenon affects local FDR just as it affects FDR: the presence of many false hypotheses, through a low value of π_0, decreases the local FDR of all other hypotheses, making their rejections seem more reliable. Relaxing the identical distribution assumption in (10), e.g. by giving each hypothesis its own π_0, would solve the cheating problem, but this relaxation makes the model unidentifiable without substantial external information.

The discussion on the role of dependence in local FDR estimation is similar to the one for Storey's methods. On the one hand, the model (10) only specifies the marginal distribution for each gene, and estimates based on this marginal model are unbiased whatever the joint distribution of test statistics [75]. On the other hand, estimates of π_0 and $f(t)$ may be highly variable if test statistics are correlated [77], and reporting a point estimate without any assessment of variability can too easily give a false sense of security. There is much recent work studying the behavior of local FDR estimates in the presence of correlation. It has been noticed that the observed distribution of test statistics of true hypotheses can be much wider or narrower than suggested by the theoretical distribution $f_0(t)$, and several authors [78, 79, 80] have suggested to use an *empirical null distribution* instead, estimated from the data. Recently, Efron [80] has worked on estimating the variance of local FDR estimates in the presence of correlations, obtaining interesting results for normally distributed test statistics. More general results are not yet available, and we await further developments in this area.

The `locfdr` software in `R` for local FDR requires z-scores that are normally distributed for true null hypotheses. Since the p-values were based on two-sided tests, we transformed the p-values of the Van de Vijver and Rosenwald data back to z-scores using

$$z_i = \begin{cases} \Phi^{-1}(\frac{1}{2}p_i) & \text{if } s_i = 1 \\ \Phi^{-1}(1-\frac{1}{2}p_i) & \text{otherwise} \end{cases},$$

where s_i is the sign of the estimated log hazard ratio for probe i, and Φ is the normal cdf. In the Van de Vijver data, Efron's method finds 1000 hypotheses with a local FDR below 0.05. The hypothesis ranked 1340th, the last hypothesis rejected by the Benjamini & Hochberg procedure, gets a local FDR estimate of 0.11. In the Rosenwald data 75 hypotheses receive local FDR below 0.05. If instead of the theoretical $\mathcal{N}(0,1)$ null distribution for the z-values the empirical null distribution was used, the results seem overly conservative. None of the local FDR values gets below 0.48 in the Van de Vijver data or 0.38 in the Rosenwald data.

6.3. FDP confidence

Point estimates of FDP are of limited value if no assessment of variability is available. For assessing such variability, a confidence interval is most informative. Goeman and Solari [11] developed such confidence intervals under the PDS assumption and for general dependence between p-values, and Meinshausen [81] developed a related permutation-based method (Section 6.4). It is interesting to note that both method are purely frequentist and do not require the additional empirical Bayes assumptions that Storey's and especially Efron's point estimation methods require.

In the approach of Goeman and Solari, the user chooses a set $\mathcal{R}$ freely, and uses the method to obtain a confidence interval for the FDP $Q(\mathcal{R})$ of this set. We make the dependence of Q on the set $\mathcal{R}$ explicit in the notation. The confidence intervals are one-sided, using the trivial lower bound $Q(\mathcal{R}) \geq 0$, because only the upper bound to the number of false dicoveries is of interest. The resulting confidence interval $[0, \bar{Q}(\mathcal{R})]$ is valid if we have

$$\mathrm{P}(Q(\mathcal{R}) \leq \bar{Q}(\mathcal{R})) \geq 1-\alpha. \tag{12}$$

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 (12) 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. The user can therefore look ate the confidence bounds $Q(\mathcal{R})$ of many sets $\mathcal{R}$ of potential interest and pick the set with the best bound, without destroying the validity of the confidence statement of the selected set. Since the method additionally 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.

All these confidence statements for $Q(\mathcal{R})$ can be alternatively presented as confidence statements on the number of false rejections $V(\mathcal{R}) = \#(\mathcal{R} \cap \mathcal{T})$ by simply multiplying by the fixed quantity $\#\mathcal{R}$ before and after the first inequality sign in (12).

Two variants of the method of Goeman and Solari are relevant for this overview: one that assumes the PDS condition, and one that holds for general dependence structures of p-values. The PDS-based method uses exactly

the same combination of the closed testing procedure [47] 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, simultaneous confidence bounds of the form (12) are obtained 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 PDS on the true hypotheses underlying the Simes-based method of Goeman and Solari is identical to the assumption underlying Hommel's and Hochberg's methods, and almost identical to that underlying the Benjamini & Hochberg procedure.

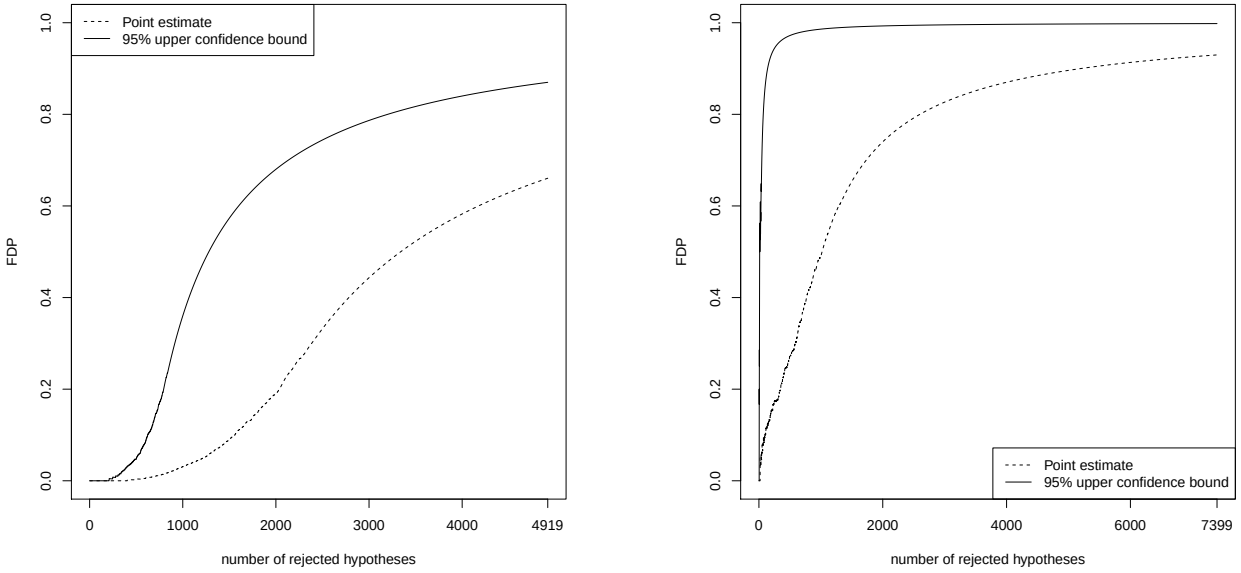
The result of the method is not a single rejected set, such as the methods from Sections 4 and 5, but rather $2^m - 1$ simultaneous confidence bounds, one for every possible subset $\mathcal{R}$ of $\mathcal{H}$. In the Van de Vijver data, taking $\mathcal{R} = \mathcal{H}$, we find that with 95% confidence there are at least 640 false hypotheses among the 4919. The smallest set containing at least 640 false 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 conservative point estimate is robust to selection of $\mathcal{R}$. The estimate has the property that it overestimates the true value at most 50% of the time, even for the final selected $\mathcal{R}$. In practice, the point estimate tends to overestimate FDP. 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. Figure 4 plots the point estimate and the 95% upper confidence bound for FDP for rejected sets consisting of the top k p-values in the Rosenwald and Van de Vijver data.

If the PDS assumption cannot be assumend, a variant of the FDP confidence bound method is available based on Hommel's inequality (2). This reduces all critical values of the Simes inequality by a factor $\sum_{k=1}^{m} 1/k$, but 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 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 confidence intervals rather than a point estimates. It also takes into account post hoc selection of $\mathcal{R}$, by making the FDP confidence



(a) Van de Vijver data

(b) Rosenwald data

Figure 4. Simultaneous upper confidence bounds and estimates of FDP for rejected sets consisting of the k smallest p-values, with k ranging from 1 to m.

intervals simultaneous for all $\mathcal{R}$. These two properties make the method suitable for the reversal of roles described in the beginning of Section 6. 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 $\bar{Q}$. The validity of the final FDP confidence interval $[0, \bar{Q}]$ 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 [82], but since these do not take the form of a single value, but of a whole confidence distribution, they are less straightforward to use.

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

6.4. Meinshausen's permutation method

Permutation-based confidence intervals for FDP of the form (12) that are analogous to those of Goeman and Solari can be obtained using a method due to Meinshausen [81]. Meinshausen's confidence bounds were initially only for rejected sets of the form $\mathcal{R} = \{H_i : p_i \leq t\}$, but 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 $\mathcal{R}$ by embedding Meinshausen's procedure in a closed testing procedure.

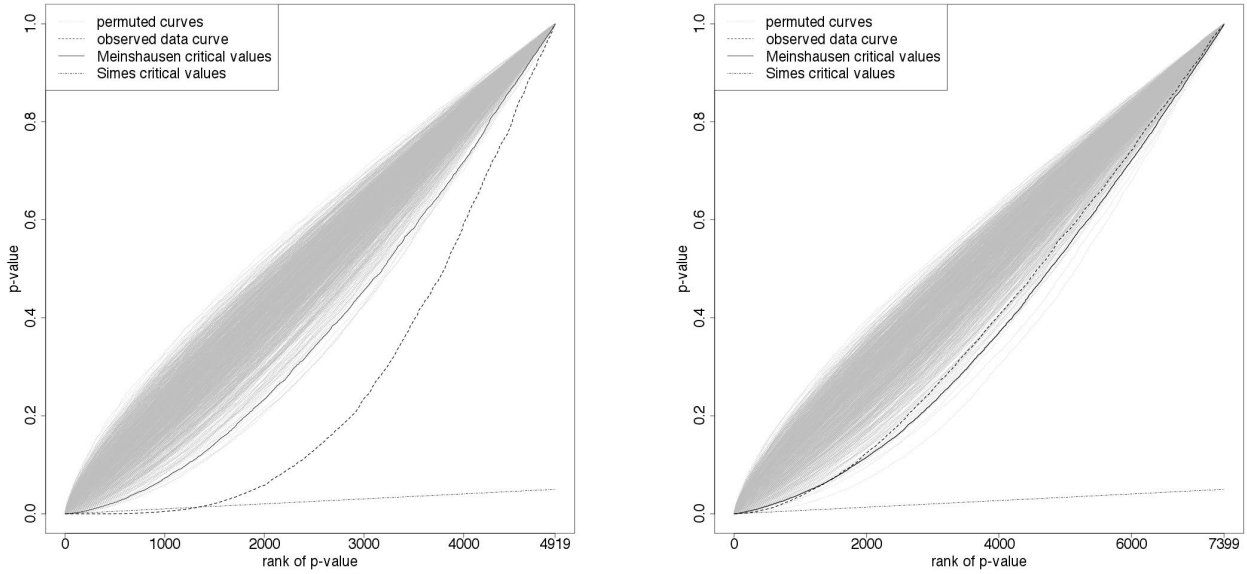
Meinshausen's method is based on a permutation-based variant of Simes' inequality, finding critical values $c_1, \ldots, c_m$ such that with probability at least $1 - \alpha$, we have simultaneously

$$q_{(i)} \geq c_i \qquad \text{for all } i = 1, \ldots, m_0, \tag{13}$$

where $q_{(1)} \leq \ldots \leq q_{(m_0)}$ are the m_0 ordered p-values of hypotheses corresponding to true hypotheses. Rather than obtaining these critical values on theoretical grounds as Simes did, Meinhausen finds them by permutation arguments. Constructing a matrix $\mathbf{P}$ of permutation p-values as in Westfall & Young's max$T$ (Section 4.4), each permuted data set provides a ranked p-value curve 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 (12).

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 5. 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.



(a) Van de Vijver data

(b) Rosenwald data

Figure 5. 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.

As a permutation method, Meinshausen's method shares with the method of Westfall && Young many advantages and disadvantages of permutation testing. Like the max$T$ method, Meinshausen's method is robust against p-values that are not marginally properly uniform. Meinshausen's method also shares the advantages of the max$T$ method described for discrete test statistics in Section 4.4. To use Meinshausen's method, valid permutations must be available, as described in Section 3. Like with Westfall & Young's max$T$, only $1/\alpha$ permutations are theoretically sufficient for the procedure to be applied, but a larger number of permutations is recommended 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 6.3, but bias-corrected point estimated are also available [83].

Meinshausen's method makes the same type of simultaneous confidence statements as Goeman and Solari's methods in Section 6.3, 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`.

6.5. 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.

Point estimates for FDP or (local) FDR are available using the methods of Storey, Efron, Goeman & Solari and Meinshausen. These can be used to examine the proportion of false positives in a rejected set. However, such estimates can be no more than rough indicators without some assessment of variability. To assess this variability, confidence intervals are preferable to standard error estimates because FDP has a strong tendency toward skewness.

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 5.4 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 5.4), 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.

7. 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 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.

7.1. 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 5.4, 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, there is always 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-value if the hypothesis is true, but precisely what can or should be used depends on the specifics of model and data [84, 85, 86, 87]. In sequencing data, genes with low median coverage will hardly have any power and are better discarded. 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 [84, 85].

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.

7.2. 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 [41, 42]. 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 7.1 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 [88, 89, 90, 91]. This type of methods can also be used in the situation that multiple hypotheses are considered for each probe, for example when doing pairwise differential expression between three groups of subjects. In this situation, researchers may first reject the overall null hypothesis for each probe that all these hypotheses are simultaneously true, before going down to the detailed tests of the individual hypotheses for that probe.

8. 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 6 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 [92, 93]. This is the famous *winner's curse*, which has been much discussed in the context of genome-wide association studies, where it shows up in the odds ratios of top SNPs [94].

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 the PDS assumption, 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. PDS-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. With one-sided tests, the PDS assumption is more problematic. 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 PDS-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. Bender R, Lange S. Adjusting for multiple testing–when and how? *Journal of Clinical Epidemiology* 2001; **54**(4):343–349.
9. Ioannidis J. Why most published research findings are false. *PLoS Medicine* 2005; **2**(8):e124.
10. Rothstein H, Sutton A, Borenstein M. *Publication Bias in Meta-Analysis*. Wiley Online Library, 2005.
11. Goeman J, Solari A. Multiple testing for exploratory research. *Statistical Science* 2011; **26**(4):584–597.
12. 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.
13. 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.
14. 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.
15. Westfall P, Tobias R, Wolfinger R. *Multiple comparisons and multiple tests using SAS*. SAS Publishing, 2011.
16. 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.
17. Finner H, Roters M. On the false discovery rate and expected type I errors. *Biometrical Journal* 2001; **43**(8):985–1005.
18. Storey J. The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* 2003; **31**(6):2013–2035.
19. Scott J, Berger J. An exploration of aspects of bayesian multiple testing. *Journal of Statistical Planning and Inference* 2006; **136**(7):2144–2162.
20. Hommel G. Tests of the overall hypothesis for arbitrary dependence structures. *Biometrical Journal* 1983; **25**(5):423–430.
21. 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.
22. Finner H, Dickhaus T, Roters M. Dependency and false discovery rate: asymptotics. *The Annals of Statistics* 2007; :1432–1455.
23. Sarkar S. Two-stage stepup procedures controlling FDR. *Journal of Statistical Planning and Inference* 2008; **138**(4):1072–1084.
24. Simes R. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**(3):751–754.
25. Sarkar S. On the Simes inequality and its generalization. *IMS Collections Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* 2008; **1**:231–242.
26. Rødland E. Simes' procedure is 'valid on average'. *Biometrika* 2006; **93**(3):742–746.
27. Sarkar S. Fdr-controlling stepwise procedures and their false negatives rates. *Journal of Statistical Planning and Inference* 2004; **125**(1):119–137.
28. 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.
29. Kim K, Van De Wiel M. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics* 2008; **9**(1):114.
30. Yekutieli D. Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test* 2008; **17**(3):458–460.
31. Good P. *Permutation tests*. Wiley Online Library, 2000.
32. Pesarin F. *Multivariate permutation tests: with applications in biostatistics*. Wiley Chichester, 2001.
33. Westfall PH, Troendle JF. Multiple testing with minimal assumptions. *Biometrical Journal* 2008; **50**(5):745–755.
34. Huang Y, Xu H, Calian V, Hsu J. To permute or not to permute. *Bioinformatics* 2006; **22**(18):2244–2248.
35. Goeman J, Solari A. The sequential rejection principle of familywise error control. *The Annals of Statistics* 2010; **38**(6):3782–3810.
36. 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.
37. Perneger T. What's wrong with Bonferroni adjustments. *British Medical Journal* 1998; **316**(7139):1236–1238.
38. Westfall PH, Troendle JF, Pennello G. Multiple mcnemar tests. *Biometrics* 2010; **66**(4):1185–1191.
39. Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 1967; **62**(318):626–633.
40. Rothman K. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; **1**(1):43–46.
41. 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.
42. Hummel M, Meister R, Mansmann U. Globalancova: exploration and assessment of gene group effects. *Bioinformatics* 2008; **24**(1):78–85.
43. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**(2):65–70.
44. Hochberg Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**(4):800–802.
45. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988; **75**(2):383–386.
46. Hommel G. A comparison of two modified Bonferroni procedures. *Biometrika* 1989; **76**(3):624–625.

47. Marcus R, Peritz E, Gabriel K. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**(3):655–660.
48. Westfall P, Young S. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley-Interscience, 1993.
49. 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.
50. Ge Y, Dudoit S, Speed T. Resampling-based multiple testing for microarray data analysis. *Test* 2003; **12**(1):1–77.
51. Van Wieringen W, Van De Wiel M, Van Der Vaart A. A test for partial differential expression. *Journal of the American Statistical Association* 2008; **103**(483):1039–1049.
52. 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.
53. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* 2001; **29**(4):1165–1188.
54. 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.
55. Benjamini Y, Krieger A, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 2006; **93**(3):491–507.
56. Blanchard G, Roquain E. Adaptive false discovery rate control under independence and dependence. *The Journal of Machine Learning Research* 2009; **10**:2837–2871.
57. 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.
58. Blanchard G, Roquain E. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics* 2008; **2**:963–992.
59. 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.
60. 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.
61. 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.
62. Ge Y, Sealfon S, Speed T. Some step-down procedures controlling the false discovery rate under dependence. *Statistica Sinica* 2008; **18**(3):881–904.
63. 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.
64. Troendle J. Comments on: Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test* 2008; **17**(3):456–457.
65. Goeman J, Mansmann U. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* 2008; **24**(4):537–544.
66. 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.
67. Storey J. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2002; **64**(3):479–498.
68. 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.
69. Schwartzman A, Lin X. The effect of correlation in false discovery rate estimation. *Biometrika* 2011; **98**(1):199–214.
70. Pounds S, Cheng C. Improving false discovery rate estimation. *Bioinformatics* 2004; **20**(11):1737–1745.
71. Owen A. Variance of the number of false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005; **67**(3):411–426.
72. Qiu X, Yakovlev A. Some comments on instability of false discovery rate estimation. *Journal of Bioinformatics and Computational Biology* 2006; **4**(5):1057–1068.
73. Tusher V, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 2001; **98**(9):5116–5121.
74. 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.
75. Efron B. Microarrays, empirical bayes and the two-groups model. *Statistical Science* 2008; :1–22.
76. Morris C. Comment: Microarrays, empirical bayes and the two-groups model. *Statistical Science* 2008; :34–40.
77. 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.
78. Pawitan Y, Calza S, Ploner A. Estimation of false discovery proportion under general dependence. *Bioinformatics* 2006; **22**(24):3025–3031.
79. Efron B. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* 2007; **102**(477):93–103.
80. Efron B. Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association* 2010; **105**(491):1042–1055.
81. Meinshausen N. False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics* 2006; **33**(2):227–237.
82. Goeman J, Solari A. Rejoinder. *Statistical Science* 2011; **26**(4):608–612.
83. 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.
84. 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.

85. Finos L, Salmaso L. FDR-and FWE-controlling methods using data-driven weights. *Journal of Statistical Planning and Inference* 2007; **137**(12):3859–3870.
86. Hackstadt A, Hess A. Filtering for increased power for microarray data analysis. *BMC Bioinformatics* 2009; **10**(1):11.
87. van Iterson M, Boer J, Menezes R. Filtering, FDR and power. *BMC Bioinformatics* 2010; **11**(1):450.
88. Meinshausen N. Hierarchical testing of variable importance. *Biometrika* 2008; **95**(2):265–278.
89. Yekutieli D. Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association* 2008; **103**(481):309–316.
90. Benjamini Y, Bogomolov M. Adjusting for selection bias in testing multiple families of hypotheses. *Arxiv preprint arXiv:1106.3670* 2011; .
91. 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.
92. Crager M. Gene identification using true discovery rate degree of association sets and estimates corrected for regression to the mean. *Statistics in medicine* 2010; **29**(1):33–45.
93. Efron B. Tweedies formula and selection bias. *Journal of the American Statistical Association* 2011; **106**(496):1602–1614.
94. 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.