PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2009) **367**, 1–17
doi:10.1098/rsta.2009.0127

Selective inference in complex research

By Yoav Benjamini[1,*], Ruth heller[2] and Daniel Yekutieli[1]

[1]*Department of Statistics and Operations Research, Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv, Israel*
[2]*Faculty of Industrial Engineering and Management, Technicon–Israel Institute of Technology, Haifa, Israel*

We explain the problem of selective inference in complex research using a recently published study: a replicability study of the associations in order to reveal and establish risk loci for type 2 diabetes. The false discovery rate approach to such problems will be reviewed, and we further address two problems: (i) setting confidence intervals on the size of the risk at the selected locations and (ii) selecting the replicable results.

Keywords: false discovery rate; false coverage rate; multiple comparisons; replicability; genome-wise association scan

1. Introduction

In the current decade we have witnessed an explosion in the size of a typical study, in terms of both the amount of data and the many research questions that fall into the domain of the study and determine its outcome. Instead of a study trying to establish associations between a disease and a few genetic markers on the genome, the current practice is to engage in a genome-wise scan for associations (GWA) involving hundreds of thousands of markers of location in the form of single nucleotide polymorphisms (SNPs) in order to identify risk loci. The question we address is how to draw inference from the few findings selected from the many tested. Moreover, there are new demands in the process of comparing one's own results with those of previous studies. The ease with which the data of completed studies can be stored has driven research funds and journals to require that even these large datasets be made publicly available. As a result, the question of replicability of results has gained a more concrete dimension: not only a verbal discussion, but a possibility to analyse one's own data together with previously collected data.

In order to make our discussion more concrete, we take a study that reports replicable results in an important area as an example. While generally the study benefits from careful and competent data analysis, issues regarding the effect of selection on the inferences made have not been addressed adequately, as we shall explain below. This is not a limitation unique to the example study. Many large and complex studies suffer from similar problems, partly because some of the statistical challenges raised by selective inference have only recently surfaced.

*Author for correspondence (ybenja@tau.ac.il).

One contribution of 11 to a Theme Issue 'Statistical challenges of high-dimensional data'.

Table 1. Odds ratio estimates, 0.95 confidence intervals (CIs) and 0.05 FCR-adjusted CIs for confirmed T2D susceptibility variants in Zeggini *et al.* (2007).

region	odds ratio	0.95 CIs	FCR-adjusted CIs
FTO	1.17	[1.12, 1.22]	[1.05, 1.30]
CDKAL1	1.12	[1.08, 1.16]	[1.03, 1.22]
HHEX	1.13	[1.08, 1.17]	[1.02, 1.25]
CDKN2B	1.20	[1.14, 1.25]	[1.07, 1.34]
CDKN2B	1.12	[1.07, 1.17]	[1.00, 1.25]
IGF2BP2	1.14	[1.11, 1.18]	[1.06, 1.23]
SLC30A8	1.12	[1.07, 1.16]	[1.01, 1.24]
TCF7L2	1.37	[1.31, 1.43]	[1.23, 1.53]
KCNJ11	1.14	[1.10, 1.19]	[1.03, 1.26]
PPARG	1.14	[1.08, 1.20]	[1.00, 1.30]

(*a*) *The type 2 diabetes study*

In the study 'Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes' by Zeggini *et al.* (2007), a UK sample of type 2 diabetes (T2D) was compared with a control sample to reveal risk loci (following the authors, we refer to it as the WTCCC1 study). The results were combined with the results of three other studies: another one from the UK and two others conducted previously by different organizations elsewhere (WTCCC2, DGI and the Fusion studies), with the intention to reveal replicated associations. Ten regions were identified as each having a large association as measured by the odds ratio. The estimates of the odds ratios and their 95 per cent confidence intervals, along with the p-values, estimated from different combinations of studies, are reported by Zeggini *et al.* (replicated in part here in table 1). The p-values of the associations reported for the WTCCC1 study range from 10^{-2} to 10^{-8}. In their combined re-analysis of the four studies, the reported p-values range from 10^{-6} to 10^{-48}. No explicit adjustment was made in the new study or in the combined analysis, neither for the fact that some 400 000 SNPs were scanned and only the selected few are reported, nor for the fact that the search for signals took place over four studies. Instead, 10^{-5} was used as a first cut-off, and a heuristic approach to the prioritization of signals was developed using the information from other studies to justify picking weaker signals in the new study. The heuristic approach included the following:

(i) p-values between 10^{-2} and 10^{-5} in the primary GWA scan,
(ii) corroborating evidence for association with T2D in the companion Diabetes Genetics Initiative (DGI) and Fusion scans,
(iii) biological candidacy of the gene, and
(iv) identification of multiple independent associations within the same locus (defined as $r^2 < 0.4$).

As the authors emphasize, 'These criteria would have led to selection of SNPs within both KCNJ11 and PPARG for second-wave replication despite the modest evidence for association based on the original WTCCC scan. Indeed, despite

concerns that differences in ethnic origin, ascertainment schemes, genotyping platforms and analysis plans across the three studies would result in effect size heterogeneity, the enhanced signals observed at known susceptibility variants in KCNJ11 ($p = 0.0013$ in WTCCC, $p \approx 5.0 \times 10^{-11}$ in combined analysis of all studies) and PPARG ($p = 0.0013$ in WTCCC, $p \approx 1.7 \times 10^{-6}$ in combined analysis) provided encouragement that this approach would highlight additional loci with high prior odds of association'.

There is quite a strong indication for some of the signals, but we cannot quantify the uncertainty involved, as follows:

(i) What does $10^{-2} < p < 10^{-5}$ mean? With 400 000 tests, we expect about 4000 tests to reach significance at the 10^{-2} level, 400 at the 10^{-3} level and four at the 10^{-5} level *if no association is true*.
(ii) Do the 10 reported 95 per cent confidence intervals still have 0.95 probability to cover their respective odds ratios?
(iii) Do p-values smaller than 10^{-6} in the combined analysis of the four experiments indicate replicated signals? In what sense?

(*b*) *Our goal*

It is quite acceptable these days in genomic research to address question (i), where a few significant results are selected out of a large pool of p-values, with the aid of statistical methods that address the issue of multiplicity. We review the false discovery rate (FDR) approach to this question, presenting the concept, the methods and variations on these concepts and methods. We then address the other two selective inference questions. The question about replicability has surfaced only recently, with the rise of complex research that combines the available datasets, and will become more and more essential as large consortia that conduct this kind of selective meta-analysis are created. We offer a coherent way of looking at this question, presenting ways to select discoveries with a lower or higher level of replicability. The setting of confidence intervals on the selected set has not been recognized until recently as a serious problem, but this is changing as the number selected, relative to the size of the pool over which the selection is conducted in a typical study, decreases. We discuss the false coverage rate approach to this problem. For all three questions, the readers can take away methods that can be immediately and directly implemented in their studies, even if better ones will be found in the future.

2. The false discovery rate approach

(*a*) *The false discovery rate*

The FDR was suggested by Benjamini & Hochberg (1995, hereafter BH) as an appropriate intermediate way of controlling for the inflated type I error in large studies. If one uses regular α-level testing in order to tag R discoveries (where a discovery is a rejected null hypothesis) among the m tested, then the number of false discoveries being made, denoted by V, may become very high. In fact, if only a few of the tested potential discoveries are true discoveries (i.e. false null hypotheses), m_1, say, and the other $m_0 = m - m_1$ are false ones (i.e. true null

hypotheses), the expected number of false discoveries is $\alpha m_0 \approx \alpha m$. With 400 000 potential discoveries of associations in our example, working at the usual $\alpha = 0.05$ results in $E(V) \approx 20\,000$ even if m_1 is in the thousands.

It is therefore clear that we need to use a stricter measure for declaring significance in such studies. The trusted Bonferroni method, in which we use $\alpha_{\mathrm{BON}} = \alpha/m$ in order to call a discovery statistically significant, is at the other extreme. It always assures that $E(V) \leq \alpha$, and therefore also that the probability of making even one error, namely the family-wise error rate (FWER), is not larger than α, but the bar for making a call about a discovery is set very high: it is $0.05/400\,000 = 1.25 \times 10^{-7}$ in our example.

The FDR takes a middle way by addressing the proportion of the false discoveries among the discoveries. That is, if R discoveries are made, define the proportion of false discoveries Q to be V/R if $R > 0$, and otherwise 0. Then the FDR is the expected proportion of false discoveries

$$\mathrm{FDR} = E(Q) = E\left[\left(\frac{V}{R}\right) I(R > 0)\right],$$

where $I(\cdot)$ is the indicator function. We advocate to control the FDR at a predefined level q in large studies.

What makes the FDR an attractive target in such studies? Screening $m = 100$ potential discoveries, making three false ones among 60 discovered is bearable; making three false ones among four discovered is unbearable. So *the FDR is adaptive* to the problem being faced, allowing more errors when the problem offers more discoveries. Moreover, the same argument holds when inspecting 100 000 findings rather than 100, so *the FDR is scalable* to the size of the problem. Suppose that nothing is real in the findings we screen, and we ensure that $\mathrm{FDR} \leq q$; it then follows that the FWER is also bounded by the same q (and hence the tendency to work with the traditional $q = 0.05$). However, if some findings are real, controlling the FDR instead of the FWER allows more discoveries.

Finally, in screening studies such as the study we use as an example, the discoveries are not the end of a story but rather the beginning of many. Each discovery is followed up with more bioinformatic and 'wet-lab' efforts, with high investment of time and money. The FDR naturally captures this effort, by controlling

$$\mathrm{FDR} = E\left(\frac{\text{efforts wasted on chasing red herrings}}{\text{total follow-up efforts}}\right). \tag{2.1}$$

From this perspective, one may sometimes choose to work with q's higher than the traditional 0.05. See Benjamini & Yekutieli (2005*a*) for a discussion of this point that offers some guidelines in genomic research.

(*b*) *Controlling the false discovery rate*

The following method was shown in BH to control the FDR at or below the desired level q under some conditions. Each hypothesis, as to whether a specific finding is real or not, is tested and a p-value is calculated. Let $H_{0i}, i = 1, \ldots, m$, be the hypotheses, and $p_i, i = 1, \ldots, m$, be the corresponding p-values. Sort the p-values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ and let $p_{(k)}$ be the largest value,

so that $p_{(k)} \le kq/m$. If no such k exists select no discovery as real. Otherwise, reject the k hypotheses corresponding to $p_{(1)}, \ldots, p_{(k)}$, declaring these findings to be discoveries.

An alternative way of presenting the results of this procedure is by presenting the adjusted p-values. The BH-adjusted p-values are defined as

$$P_{(i)}^{\mathrm{BH}} = \min\left(\left(\min_{j \ge i} mp_{(j)}/j\right), 1\right).$$

Then $P_{(i)}^{\mathrm{BH}} \le q$ if and only if $H_{(i)}$ is among the discoveries when using the procedure in BH at level q.

(*c*) *Controlling the false discovery rate in the type 2 diabetes study*

We demonstrate the procedure in BH using the WTCCC1 study. The first column of table 2 carries the information about the p-values. Let us assume for the sake of this demonstration that these are the most extreme 10 among the approximately 400 000 over which the scan was made. Sorting them we get the 10 sorted p-values: 6.7×10^{-13}, 2.0×10^{-8}, 5.4×10^{-6}, 2.5×10^{-5}, 3.2×10^{-4}, 7.6×10^{-4}, 1.7×10^{-3}, 1.3×10^{-3}, 1.3×10^{-3}, 2×10^{-2}. Multiplying the i^{th} p-value by 400 000 and dividing by i, we get the sequence $2.7 \times 10^{-7}, 4.0 \times 10^{-3}, 7.2 \times 10^{-1}, 2.5, \ldots$, the other six also being bigger than 1. Taking for each i, the minimum overall $p_{(j)}$, such that $j \ge i$ and 1, we get the above three as the only BH-adjusted p-values that are smaller than 1, all others being 1. At the conventional FDR level of 0.05 only two discoveries have been made. For the DGI study using similar analysis, we get four BH-adjusted p-values that are smaller than 1, and all these are also smaller than 0.01. Thus, in this study, all four are statistically significant at any reasonable level.

(*d*) *Validity under dependency*

If the test statistics from which the p-values are calculated are independent, or positively dependent, the FDR is controlled by the above method at level $qm_0/m \le q$. For the exact conditions, see BH and Benjamini & Yekutieli (2001). In particular, the condition for positive dependency addresses block dependencies that result very naturally in association studies because of haplotypes. It was also shown in the latter that, if the procedure in BH is used with $q/(1 + 1/2 + \cdots + 1/m)$, the FDR is always controlled at level $qm_0/m \le q$. This result has sometimes been interpreted as a warning that the method without the inflating factor does not control the FDR at level q. This is definitely not the case, as many simulation studies by other authors have shown.

The joint p-value distribution in which the FDR of the BH procedure reaches its upper bound (Guo & Rao 2008) is an extremely unusual distribution. For example, under the complete null hypothesis, the following holds: for $j = 1, \ldots, m$, with probability q/j, a randomly drawn subset of j p-values are set precisely in the interval $[q \times (j-1)/m, q \times j/m]$, while the remaining $m - j$ p-values are set in $[q, 1]$. In this unusual case $P_{(j)}^{\mathrm{BH}}$ and also the other FDR methods described in the following sections fail to offer the right protection without the correction factor.

Table 2. The p-values in four cohorts that correspond to the genes judged to be associated with T2D in Zeggini *et al.* (2007), as well as the Fisher-based partial conjunction p-values for $u = 1, \ldots, 4$. n.a., not applicable.

gene ID	WTCCC1	WTCCC2	DGI	Fusion	$p^{1/n}(g)$	$p^{2/n}(g)$	$p^{3/n}(g)$	$p^{4/n}(g)$
FTO	2.0×10^{-8}	5.4×10^{-7}	0.25	0.017	4.4×10^{-13}	5.0×10^{-7}	0.0275	0.25
CDKAL1	2.5×10^{-5}	8.3×10^{-5}	2.4×10^{-3}	9.5×10^{-3}	2.5×10^{-10}	4.2×10^{-7}	2.6×10^{-4}	9.5×10^{-3}
HHEX	5.4×10^{-6}	0.02	1.7×10^{-4}	0.025	1.9×10^{-9}	1.2×10^{-5}	0.0043	0.025
CDKN2B	7.6×10^{-4}	1.7×10^{-4}	5.4×10^{-8}	2.2×10^{-3}	1.6×10^{-13}	7.5×10^{-8}	2.4×10^{-5}	2.2×10^{-3}
CDKN2B	3.2×10^{-4}	8.6×10^{-4}	0.5	0.039	7.3×10^{-6}	0.0012	0.096	0.5
IGF2BP2	1.7×10^{-3}	0.018	1.7×10^{-9}	2.4×10^{-4}	1.3×10^{-13}	1.4×10^{-6}	3.5×10^{-4}	0.018
SLC30A8	0.02	1.2×10^{-3}	0.047	7×10^{-5}	1.9×10^{-7}	1.2×10^{-4}	0.0075	0.047
TCF7L2	6.7×10^{-13}	—	2.3×10^{-31}	1.4×10^{-8}	0	0	1.4×10^{-8}	n.a.
KCNJ11	1.3×10^{-3}	—	1×10^{-7}	0.014	7.2×10^{-10}	2.2×10^{-4}	0.014	n.a.
PPARG	1.3×10^{-3}	—	0.019	1.4×10^{-3}	5.7×10^{-6}	3.1×10^{-4}	0.019	n.a.

On the other hand, Reiner-Benaim (2007) showed, via a combination of simulations and analytical results that, for *normally* distributed two-sided test statistics under any correlation structure,

$$\mathrm{FDR} \leq \frac{qm_0}{m} \times \left(1 + \frac{m_0}{m}(1 - m_0/m)\right) \leq q. \tag{2.2}$$

Since log odds ratios are approximately normally distributed, the method in BH can be used to analyse the data from the T2D study in spite of the possible dependencies.

(*e*) *Other methods that control the false discovery rate*

When using the procedure in BH, we actually control the FDR at a level lower than what we are willing to accept, i.e. qm_0/m instead of q. It is natural to try to estimate m_0, or equivalently the factor $p_0 = m_0/m$. Numerous efforts have been made in this direction, both theoretical and empirical, plugging the estimated p_0 back into the procedure in BH, by replacing q by qm/m_0. However, such methods introduce more variability; hence assuring theoretically the FDR control of such procedures is not easy to establish. Methods of this nature that have proven control of the FDR, at least under independence, are the one by Storey *et al.* (2003), others by Blanchard & Roquain (in press), the two-stage procedure by Benjamini *et al.* (2006) and the multiple-stage one in Gavrilov *et al.* (2009). The last two have also been shown via simulation studies to have quite tight control of the FDR even in situations with strong positive dependency (Romano *et al.* 2008).

Nevertheless, in complex research as in the example we use here, the number of potential discoveries will usually be very small relative to the total number, because the discoveries are not likely to be important if abundant in the extremely large pool searched. With $p_0 \approx 1$ there is no advantage in estimating it, and in fact the performance actually deteriorates because of the extra protection needed to combat the estimation error. Hence, we shall not dwell further here on the details of these procedures.

(*f*) *Other approaches to the false discovery rate*

Two variations on the concept of FDR were discussed in BH, before adopting the version we have presented. The first involves conditioning on making any discovery, $p\mathrm{FDR} = E(V/R|R > 0)$, later emphasized as the positive false discovery rate in Storey (2002); the other one is the ratio of expectations, $Fdr = E(V)/E(R)$. Much research has been devoted to these concepts because they are natural in the Bayes and empirical Bayes frameworks. They are also often encountered in the bioinformatic and biostatistical literature about the analysis of microarray experiments, where they are presented within the framework of a mixture model. In this model, each hypothesis about a gene's expression has $p_0 < 1$ probability of being true with the distribution of the corresponding p-value being uniform $F_0 \sim U(0, 1)$, and probability $1 - p_0$ of being false with an unknown continuous distribution F_1. The practical differences among the three concepts in large problems are not big, as we get $\mathrm{FDR} = p\mathrm{FDR} = Fdr$ in the asymptotic version of the model, where the number of genes m tends to ∞.

However, with $p_0 = 1$ being a real possibility, meaning there is no true result to be discovered in the pool over which we search, only the control of FDR offers a valid approach.

Efron and co-authors have expanded the family of FDR concepts further into the Bayesian and empirical Bayesian territory by working with the local FDR, $fdr(z) = p_0 f_0(z)/f(z)$, and offering methods that incorporate the estimation of the density under the null $f_0(z)$, of the density under the mixture of the null with the alternative $f(z)$, and the estimation of p_0. For a very interesting and accessible review of these ideas, see the text and discussion of Efron (2008).

Yekutieli (2009) explains the role of selection in controlling the occurrence of false discoveries in Bayesian analysis, and argues that selection may also affect Bayesian inference, and especially Bayesian inference based on subjective priors. He further introduces selection-adjusted Bayesian methodology based on the conditional posterior distribution of the parameters given selection. In particular, he shows that the local FDR and positive FDR can be expressed as selection-adjusted Bayesian inference for the two-group mixture model. The above approaches using the mixture model essentially cope with the selection problem by avoiding a subjective p_0 in the model via the empirical estimation of p_0.

As a final note, while differences exist, the common theme of all the efforts discussed in this section is that the issue of selective inference should be addressed. Just as we formally address the uncertainty involved in estimation or in the testing of a single hypothesis, with the aid of well-defined statistical methods, so should we make use of a fully specified method in order to cope with the problems posed by selective inference.

3. Confidence intervals for selected parameters

(*a*) *The coverage problem*

Too often the decisions as to what clues to follow with expensive research are based on significance only, rather than on estimated effect sizes. Our example study takes the better route, reporting the estimated odds ratio and its confidence interval associated with each of the few selected locations on the genome. They follow the practice of reporting the intervals of the selected set at their marginal (nominal, unadjusted) level. Unfortunately, it is common practice to ignore multiplicity when it comes to multiple confidence intervals, even though adjusting the testing of hypotheses for multiplicity has become widespread (but not mandatory as our example testifies).

Benjamini & Yekutieli (2005*b*) demonstrate that confidence intervals constructed for selected parameters cannot ensure nominal coverage probability. They suggest the false coverage-statement rate (FCR) as the appropriate criterion for capturing the error for confidence intervals constructed for selected parameters. The false coverage rate is defined in a way that parallels the definition of the FDR. Let R now be the number of confidence intervals constructed, and let V be the number of confidence intervals that do not manage to cover their respective parameters. Then V/R is the proportion of the intervals made that fail to cover (where again $V/R = 0$ when $R = 0$), and $\mathrm{FCR} = E[(V/R)I(R > 0)]$.

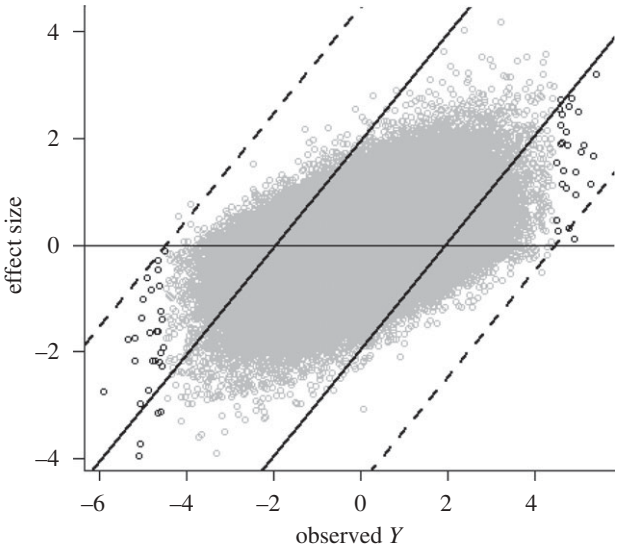

Figure 1. Simulated example—scatter plot of (y_i, θ_i). The observations shown in black are the 58 level 0.05 BH procedure discoveries. The solid lines are the 0.95 confidence intervals. The dashed lines are the 0.05 FCR-adjusted confidence intervals.

Note that, when requiring that FCR $\leq q$, we do not require that the confidence intervals have simultaneous coverage of $1-q$, a much stronger requirement. We merely require coverage-on-the-average, but the average is taken over the selected intervals only.

Figure 1 displays a simulated example of 400 000 realizations of (θ_i, Y_i), where θ_i are i.i.d. receiving the values $\pm\exp(3)$ with equal probability and $Y_i|\theta_i \sim N(\theta_i, 1)$. One can consider θ_i as the association log-odds ratio and Y_i as its estimator. The observations shown in black are the $R=58$ discoveries produced by the level 0.05 BH procedure, applied to the two-sided p-values $p_i = 2\times\{1-\Phi(|Y_i|)\}$; the remaining observations are shown in grey. The solid lines are the marginal 0.95 confidence intervals $Y_i \pm Z_{1-0.05/2}$; they cover 0.949 of all 400 000 θ_i realizations, but only six of the 58 BH discoveries; thus $V/R = 0.90$.

(*b*) *Constructing false discovery rate-controlling confidence intervals*

Benjamini and Yekutieli (2005*b*) set the following framework for discussing the false coverage rate. They assume that there are m parameters $\theta_1, \ldots, \theta_m$, with corresponding estimators $T_1, \ldots, T_m$, and the goal is to construct valid confidence intervals for the parameters selected by a given selection criterion that may depend on the value that the estimators can take, $\tilde{S}(T_1, \ldots, T_m) \subseteq \{1, \ldots, m\}$. Let R be the number of selected parameters, i.e. $R = |\tilde{S}|$. As a general method for ensuring FCR $\leq q$ for any selection criterion, they suggest constructing marginal $1 - R\times q/m$ confidence intervals for each of the R selected parameters.

When using the procedure in BH to select the parameters of interest, by way of testing whether each θ_i equals its null value θ_i^0 versus the alternative $\theta_i \neq \theta_i^0$, as demonstrated here for the T2D study (where each odds ratio was tested as to whether or not it is 1), we get a very desirable property: for each selected θ_i, the level-q FCR-adjusted confidence interval will not cover θ_i^0. This property does not always hold, as it also relies on the fact that the confidence intervals for the log-odds ratios are symmetric about the estimated log-odds ratios. Whenever this fact holds, so does the desirable connection between FDR testing using BH and FCR confidence intervals.

Figure 1 further displays the FCR-adjusted confidence intervals. The dashed lines are the 0.05 FCR-adjusted confidence intervals $Y_i \pm Z_{1-R\times 0.05/(2\times 400\,000)}$; they cover 57 of the 58 θ_i and yield $V/R = 0.017$. Note that the remedy offered by the FCR-adjusted confidence intervals, as well as the lack of coverage by the standard 95 per cent confidence intervals, persists even when 'no hypothesis is true', in the sense that $\theta_i \neq \theta_i^0$ for all i.

(c) *Selected confidence intervals in the type 2 diabetes study*

Zeggini *et al.* (2007) highlighted 10 locations associated with T2D. Odds ratio estimates based on the joint analysis of the four studies, as well as 0.95 confidence intervals, are given there. Table 1 reproduces these results in the first three columns. Since 10 parameters are chosen out of 400 000, the 0.05 FCR-adjusted confidence intervals are the marginal $1 - 10 \times 0.05/400\,000$ intervals, given in the last column of the table. They are expected to offer approximately by 0.95 coverage probability for each selected odds ratio. None of the FCR-adjusted odds ratio confidence intervals covers 1, though two are close enough to be indistinguishable.

4. Selecting replicable results

(a) *The conjunction approach to replicability*

Consider the search for replicable results across n studies. The approach taken in the T2D study is to combine the data of a single location across the studies, and analyse them jointly as if they were a single meta-study. In principle, we end up with m analyses for m locations, each summarized by a p-value and an estimate of the size of the effect. From these the best results are selected. The results of this meta-study regarding all 10 locations on the genome reported in Zeggini *et al.* (2007, table 1) were found to be statistically significant even after adjusting for the selection effect using the procedure in BH.

Nevertheless, this fact does not imply that the findings are replicable. The strongest scientific statement about the replicability of a discovery would be to establish that the discovery was repeated in all the studies examining it, even though the studies might have differed in terms of the populations (different cohorts), the diagnostic methods, the laboratory methods or the statistical methods used. This is at the heart of the experimental scientific dogma: two different experiments showing the same result, each having 1000 cases, are a better evidence than a single experiment with 2000 cases showing the same result, because they offer evidence that the result is replicable.

(*b*) *The partial conjunction approach to replicability*

The *partial conjunction hypothesis*, introduced in Benjamini & Heller (2008), sets a framework for discussing replicability. With m research questions all studied in each of n studies, we define the null hypothesis that discovery regarding g is not true in study j to be $H_{0j}(g)$. If we try to show that the discovery is true in at least u studies, with $1 \le u \le n$, we may test the complementary statement as our null hypotheses: test whether at most $n-u+1$ individual null hypotheses are true. We can phrase this hypothesis differently: let k be the (unknown) number of true discoveries (false null hypotheses); then the partial conjunction null is written as $H^{u/n}(g) : k < u$. Rejecting this hypothesis implies that the discovery regarding g has been replicated in at least u studies.

In the T2D study, the investigators actually tested the global null at each location—that the tested hypotheses are true in all studies $\cap_{i=1}^{n} H_{0i}(g)$. This hypothesis is actually $H^{1/n}(g)$, i.e. $u=1$, meaning that rejecting this null hypothesis implies that the discovery is true in at least one study. However, it does not imply that the discovery is true in more than one study, as the global null may be rejected just because the finding was extremely significant in only one study. For example, in the T2D study small p-values in the test of the global null may be due to only one cohort (or study) having highly significant associations with diabetes. Nevertheless, the association might have failed to appear in more than one study—rejecting the global null does not at all indicate a replicable discovery.

Ideally, we would have liked to show that the discovery can be observed in all studies in order to claim that it is replicable. That is, we would like to be able to reject the null hypothesis of the full conjunction with $u=n$, namely $H^{n/n}(g)$. Because of the differences in the many factors affecting the results in each particular study, this may not be feasible. It is also possible that we may simply lack power to support such a strong statement even if it is true across all studies.

Instead, we allow increasingly weaker replicability statements stating that the discoveries were real in at least all but one study, all but two studies, etc. That is, we can test the partial conjunction hypotheses $H^{u/n}(g)$, with $u=n, n-1, n-2, \ldots, 2$, offering increasingly weak replicability results. The largest u we can still reject is the strongest statement we can make.

(*c*) *Testing of partial conjunction hypotheses at a fixed u*

For a location on the genome g, let $p_{(1)}(g) \le \cdots \le p_{(n)}(g)$ be the ordered p-values from n studies. Benjamini & Heller (2008) introduced p-values for testing the partial conjunction hypothesis. We present first one version based on the Fisher test, which is valid when the p-values across studies are independent,

$$p^{u/n}(g) = P\left(\chi^2_{2(n-u+1)} \ge -2\sum_{i=u}^{n} \log p_{(i)}(g)\right), \tag{4.1}$$

where χ^2_ν denotes a chi-squared random variable on ν d.f.

For testing a large family of partial conjunction hypotheses, $H_0^{u/n}(g)$, $g=1,\ldots,m$, Benjamini & Heller (2008) suggested the following procedure: first combine the n p-values for every g as appropriate for testing $H_0^{u/n}(g)$, and

then use an FDR controlling procedure on the partial conjunction p-values $p^{u/n}(g), g=1,\ldots,m$. (The number of studies combined per location, n, may vary from one location to the next, so a more precise yet cumbersome notation for the partial conjunction p-value at g is $p^{u/n(g)}(g)$.)

Other choices of p-values for partial conjunction hypotheses that are discussed there combine the same $n+1-u$ largest p-values using Bonferroni, Simes or other tests, and the choice is guided by the joint distribution of the p-values at g.

(*d*) *Testing of partial conjunction hypotheses at all levels of u*

It is not clear as to what value of u should be chosen in order to establish replicability. In practice, instead of predefining u, it is possible to test in order the partial conjunction hypotheses with $u=1,2,\ldots,n$. The following procedure can be applied to get a lower bound on the true number of true discoveries for every g that has been rejected.

Step 1. For each gene $g, g=1,\ldots,m$ compute the global null p-value $p^{1/n}(g)$.
Step 2. Apply the BH procedure at level q on $\{p^{1/n}(g): g=1,\ldots,m\}$. Let R be the number of rejected hypotheses.
Step 3. For each of the R units where the global null hypothesis has been rejected, test sequentially the partial conjunction hypotheses $u=2,3,\ldots,n$ at level Rq/m:

$$u_{\max}(g)=\left(\arg\min_{u>1}\left\{p^{u/n}(g)>\frac{Rq}{m}\right\}\right)-1.$$

The results of this procedure can be presented in terms of the adjusted p-values as follows: for $u=1$ the ith largest adjusted p-value is $\tilde{p}^{1/n}_{(i)}=\min(\min_{j\geq i} mp^{1/n}_{(j)}/j, 1)$, and for $u>1$ the adjusted p-value is

$$\tilde{p}^{u/n}(g)=\frac{m}{\sum_{g=1}^{m} I[\tilde{p}^{1/n}(g)\leq q]}p^{u/n}(g).$$

Then, $H^{u/n}(g)$ is among the discoveries if and only if $\tilde{p}^{u/n}(g)\leq q$.

The following theorem shows that, by applying the above procedure, we expect only a fraction q of the lower bounds to be larger than the true number of studies where the findings are true.

Theorem 4.1. *If the test statistics across units $g\in\{1,\ldots,m\}$ are independent, then the expected proportion of lower bounds that exceed the number of false null hypotheses ($u_{\max}(g)>k(g)$, where $k(g)$ is the number of false null hypotheses), out of all R units rejected, is q.*

The proof is in appendix A. See Benjamini & Heller (2008) and Heller *et al.* (2009) for related results.

Table 3. For each gene, the lower bound $u_{\max}$ on the number of studies in which it was found to be associated with T2D for the procedure in §4*d* with $q = 0.05$, as well as the associated p-value and adjusted p-value. The genes are sorted first by decreasing $u_{\max}$, then by increasing $p^{u_{\max}/n}$.

gene ID	$u_{\max}$	$p^{u_{\max}/n}$	$\tilde{p}^{u_{\max}/n}$
TCF7L2	3	1.4×10^{-8}	7.0×10^{-4}
CDKN2B	2	7.5×10^{-8}	0.00375
CDKAL1	2	4.2×10^{-7}	0.021
FTO	2	5.0×10^{-7}	0.025
IGF2BP2	1	1.3×10^{-13}	5.2×10^{-8}
KCNJ11	1	7.2×10^{-10}	0.000144
HHEX	1	1.9×10^{-9}	0.00025
SLC30A8	1	1.9×10^{-7}	0.019

So far, the control over false-positive lower bounds has been established theoretically only for partial conjunction p-values that are independent across locations. Still, the results about the stability of the procedure in BH under dependency encourage us to advocate its use in settings similar to those of the T2D study. Further verification of the validity of the procedure in settings with SNP-type dependence is a direction for future research.

(*e*) *Replicable findings in the type 2 diabetes studies*

Table 2 shows in columns 2–5 the p-values in the four cohorts that correspond to the genes judged to be associated with T2D in Zeggini *et al.* (2007), and in columns 6–9 the Fisher-based partial conjunction p-values for $u = 1, \ldots, 4$, respectively. Note that the ranking of the genes based on partial conjunction p-values changes with the choice of u. Specifically, the three most significant p-values are as follows (in order): for $u = 1$, TCF7L2, IGF2BP2, CDKN2B; for $u = 2$, TCF7L2, CDKN2B, CDKAL1. As discussed above, the ranking based on $u = 2$ makes more sense than the ranking based on $u = 1$ if the goal is to establish the replicability of the finding.

Some locations were not tested in the WTCCC2 study, so we continued our analysis by conservatively assuming that the missing p-values are 1. Applying the BH procedure on the partial conjunction p-values $p^{2/n}, g = 1, \ldots, 400\,000$, at level $q = 0.05$ resulted in two discoveries: TCF7L2 and CDKN2B. Applying the BH procedure on the partial conjunction p-values $p^{3/n}, g = 1, \ldots, 400\,000$ resulted in one discovery: TCF7L2.

It is quite possible that applying the partial conjunction approach, which makes use of the p-values in the individual studies (and is therefore much easier to carry out for all), would have yielded more replicable results.

Next, we applied the procedure in §4*d* at level $q = 0.05$. The BH procedure on the p-values for testing the global null $p^{1/n}, g = 1, \ldots, 400\,000$ resulted in $R = 8$ discoveries, and subsequent tests were performed at level $\alpha = Rq/m = 8 \times 0.05/400\,000 = 10^{-6}$. Table 3 shows the following results: the association between TCF7L2 and T2D was discovered in at least three studies; for three genes an association with T2D was discovered in at least two studies; for four genes an association with T2D was discovered in at least one study.

5. Other developments and future directions

(a) Using weights

One objection to our structured approach to selective inference may be that it is too formal, and it fails to incorporate outside information other than the marginal p-values. For example, in the T2D study, the authors claim that they have also considered two other factors when making their choice: the biological candidacy of the gene and the identification of multiple independent associations within the same locus.

This possible objection is unfounded. Both Benjamini & Hochberg (1997) and Genovese *et al.* (2006) employ weights to differentiate between the hypotheses tested within the FDR framework. The weights may incorporate differing importance of the hypotheses (in the first approach), where the weighted FDR is defined as

$$\mathrm{FDR} = E\left(\frac{\sum_{i=1}^{m} w_i V_i}{\sum_{i=1}^{m} w_i R_i} I\left(\sum_{i=1}^{m} w_i R_i > 0\right)\right),$$

V_i and R_i being a false rejection and a rejection of an individual hypothesis i, respectively, and w_i its weight. In the study discussed here, the weights can reflect both the biological importance of the association between a SNP and the disease, information that is available prior to conducting the analysis, and the consistency of the association across nearby SNPs, information which is available from the analysis. The use of the second type of information does not increase the selection problem because it is statistically independent from the significance of the association at a SNP.

The weights can alternatively reflect different prospects for showing effects (in the second approach). Now the definition of the FDR remains unchanged, but the procedure is different, offering more power to reject a hypothesis that was assigned a higher weight, without an increase in the FDR. Thus, the two available studies (DGI and Fusion) can be used to define the weights, these in turn being used to explore the two UK studies.

Finally, given that some studies may be more reliable or important than others, different weights may be assigned *a priori* to these studies. A testing strategy that takes the weights into account may be more powerful and more reproducible. The partial conjunction tests will be more general than the tests introduced in §4*c*, by allowing different studies to be weighted differently when combined. Moreover, the power of the test may increase as well as the reproducibility of the finding. Specifics follow. Since the partial conjunction null hypothesis is the union of all $\binom{n}{n-u+1}$ intersection hypotheses of size $n-u+1$, $\cap_{k=1}^{n-u+1} H_{j_k}$ $\{j_1, \ldots, j_{n-u+1}\} \subset \{1, \ldots, n\}$, the weighted Fisher method (Hedges & Olkin 1985) can be used to test for partial conjunction hypotheses as follows. Let $P_w = \prod_{i=1}^{n-u+1} U_i^{w_{j_i}}$, where $U_1, \ldots, U_{n-u+1}$ are independent $U(0,1)$ random variables, and let $p_w = \prod_{i=1}^{n-u+1} p_{j_i}^{w_{j_i}}$ be the test statistic for the intersection hypothesis. The p-value is therefore $f(p_{j_1}, \ldots, p_{j_{n-u+1}}) = Pr(P_w \le p_w)$. The partial conjunction p-value is the largest of the $\binom{n}{n-u+1}$ intersection hypotheses p-values, $p^{u/n} = \max_{\{j_1,\ldots,j_{n-u+1}\} \subset \{1,\ldots,n\}} f(p_{j_1}, \ldots, p_{j_{n-u+1}})$.

It is straightforward to show that this is a valid p-value, i.e. $Pr(P^{u/n} \leq \alpha) \leq \alpha$ if $H^{u/n}$ is true. For unit weights, the partial conjunction p-values coincide with the p-values in §4*c*.

(*b*) *False discovery rate-controlled hierarchical search*

When confronted with complex research problems, with very few potential discoveries, one can benefit from a hierarchical search strategy. This can be in the form of first testing some screening hypotheses, and then focusing attention on a promising subset of the original pool of hypotheses. Alternatively, it can be achieved by collecting hypotheses in sets in which they are likely to be true together, or false together. Testing first the sets, and thereby increasing the signal-to-noise ratio, one can follow with tests within the promising sets. For the general theoretical formulation of such a selection process, see Yekutieli (2008), and for a practical implementation in a very large study involving the associations between gene expression in brain regions and measures of exploratory behaviour, see Reiner-Benaim *et al.* (2007).

In other gene-expression studies, the approach is referred to as gene-enrichment analysis where the clustering of genes into sets is based on external information regarding the pathways involved. This need not be the only way to partition the hypotheses. In brain-imaging experiments, for example, the clusters of hypotheses about brain regions, not necessarily of same size or shape, can be based on a pilot study (Benjamini & Heller 2007). The partitioning can also be based on thresholding or on a moving window (see also Pacifico *et al.* 2007).

The essence of this approach should be clear. When the tested parameters have further structure, in the sense that we have a grasp of in which sets the hypotheses are going to be true together and false together (correlated parameters), hierarchical analysis is of great potential: in many cases not only is the signal-to-noise ratio increased but the multiplicity problem can be reduced.

We should add a warning, however. The hierarchical search should be so structured as to allow control of the FDR. It is not merely enough to partition the hypotheses into sets and to control the FDR separately in each set, as is sometimes advocated.

(*c*) *Final remarks*

The selective inference issue is extremely important in large and complex studies. We have emphasized the FDR approach, which has turned out to be inherently scalable, in the sense that it has stood up to the challenges of searching for a few discoveries from among hundreds of thousands and even millions of hypotheses. The FDR approach is relevant—possibly because of its triple Frequentist/Empirical Bayes/Bayes interpretation.

We expect that the original tools developed along with the FDR approach will continue to evolve. Just as the need to address replicability came to the fore only very recently, and the effect of selection on confidence intervals was not well recognized formerly, we expect that the practical challenges in future applications will continue to shape the methodologies we offer. Still, we already have enough statistical tools to cope effectively with selective inference problems, so that a study cannot be considered valid without addressing these issues.

This research was partly funded by grants from the US National Institute of Health, the German Israel Fund and the Israel Science Foundation.

Appendix A

Proof of theorem 4.1

Let $I_0 \subset \{1, \ldots, m\}$ be the index set for units that have at least one true null hypothesis, so that $H^{n/n}$ is true. Let $k(g)$ be the true number of studies that show an effect for unit g, and let $u_{\max}(g)$ be the result of the procedure in §4*d*. Let $V(g) = 1$ if $u_{\max}(g) > k(g)$, i.e. the lower bound exceeds the true number of studies that show an effect, and 0 otherwise. Let $R(g) = 1$ if $H^{1/n}(g)$ is rejected and 0 otherwise. Let $Q = \sum_{g=1}^{m} V(g) / \sum_{g=1}^{m} R(g)$ if at least one unit g was discovered and 0 otherwise. Q is the proportion of units with false lower bounds out of all units discovered. We are interested in the quantity $E(Q)$, where

$$\begin{aligned} E(Q) &= \sum_{g=1}^{m} \sum_{l=1}^{m} \frac{1}{l} Pr\left(V(g) = 1 \cap \sum_{i=1}^{m} R(i) = l \right) \\ &= \sum_{g \in I_0} \sum_{l=1}^{m} \frac{1}{l} Pr\left(V(g) = 1 \cap \sum_{i=1, i \neq g}^{m} R(i) = l - 1 \right) \\ &= \sum_{g \in I_0} \sum_{l=1}^{m} \frac{1}{l} Pr\left(V(g) = 1 \cap C_l^{(g)} \right), \end{aligned}$$

where $C_l^{(g)}$ is the event that $p_{(l-1)}^{(g)} \leq lq/m, p_{(l)}^{(g)} > (l+1)q/m, \ldots, p_{(m-1)}^{(g)} > q$, in which $p_{(1)}^{(g)} \leq \cdots \leq p_{(m-1)}^{(g)}$ are the ordered coordinates of the vector $\vec{p}^{(g)}$ of p-values for testing the global null excluding that of g. Continuing, we have

$$\begin{aligned} E(Q) &= \sum_{g \in I_0} \sum_{l=1}^{m} \frac{1}{l} Pr\left(V(g) = 1 \cap C_l^{(g)} \right) \\ &= \sum_{g \in I_0} \sum_{l=1}^{m} \frac{1}{l} Pr\left(P^{(k(g)+1)/n}(g) \leq \frac{lq}{m} \cap C_l^{(g)} \right) \\ &= \sum_{g \in I_0} \sum_{l=1}^{m} \frac{1}{l} Pr\left(P^{(k(g)+1)/n}(g) \leq \frac{lq}{m} \right) Pr(C_l^{(g)}), \end{aligned}$$

where the last equality follows since the p-values are independent. Since $H^{(k(g)+1)/n}(g)$ is a null hypothesis, $Pr(P^{(k(g)+1)/n}(g) \leq lq/m) \leq lq/m$. The result follows:

$$E(Q) \leq \sum_{g \in I_0} \sum_{l=1}^{m} \frac{1}{l} \times \frac{lq}{m} \times Pr(C_k^{(g)}) = \sum_{g \in I_0} \frac{q}{m} \leq q.$$

■

References

Benjamini, Y. & Heller, R. 2007 False discovery rate for spatial data. *J. Am. Stat. Assoc.* **102**, 1272–1281. (doi:10.1198/016214507000000941)

Benjamini, Y. & Heller, R. 2008 Screening for partial conjunction hypotheses. *Biometrics* **64**, 1215–1222. (doi:10.1111/j.1541-0420.2007.00984.x)

Benjamini, Y. & Hochberg, Y. 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300.

Benjamini, Y. & Hochberg, Y. 1997 Multiple hypotheses testing with weights. *Scand. J. Stat.* **24**, 407–419. (doi:10.1111/1467-9469.00072)

Benjamini, Y. & Yekutieli, D. 2001 The control of the false discovery rate under dependency. *Ann. Stat.* **29**, 1165–1188.

Benjamini, Y. & Yekutieli, D 2005*a* Quantitative trait loci analysis using the false discovery rate. *Genetics* **171**, 783–789. (doi:10.1534/genetics.104.036699)

Benjamini, Y. & Yekutieli, D. 2005*b* False discovery rate-adjusted multiple confidence intervals for selected parameters *J. Am. Stat. Assoc.* **100**, 71–81. (doi:10.1198/016214504000001907)

Benjamini, Y., Krieger, A. & Yekutieli, D. 2006 Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507. (doi:10.1093/biomet/93.3.491)

Blanchard, G. & Roquain, E. In press. Adaptive FDR control under independence and dependence. *J. Mach. Learn. Res.*

Efron, B. 2008 Microarrays, empirical Bayes and the two-groups model (with discussion). *Stat. Sci.* **23**, 1–47. (doi:10.1214/07-STS236)

Gavrilov, Y., Benjamini, Y. & Sarkar, S. 2009 An adaptive step-down procedure with proven FDR control under independence. *Ann. Stat.* **37**, 619–629. (doi:10.1214/07-AOS586)

Genovese, C. R., Roeder, K. & Wasserman, L. 2006 False discovery control with p-value weighting. *Biometrika* **93**, 509–524. (doi:10.1093/biomet/93.3.509)

Guo, W. & Rao, M. B. 2008 On control of the false discovery rate under no assumption of dependency. *J. Stat. Plann. Infer.* **138**, 3176–3188. (doi:10.1016/j.jspi.2008.01.003)

Hedges, L. & Olkin, I. 1985 *Statistical methods for meta-analysis*. London, UK: Academic Press.

Heller, R., Manduchi, E., Grant, G. & Ewens W. 2009 A flexible two stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics* **25**, 1019–1205. (doi:10.1093/bioinformatics/btp076)

Pacifico, M. P., Genovese, C., Verdinelli, I. & Wasserman, L. 2007 False discovery control for random fields. *J. Multivariate Anal.* **98**, 1441–1469.

Reiner-Benaim, A. 2007 FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biometric. J.* **49**, 107–126. (doi:10.1002/bimj.200510313)

Reiner-Benaim, A., Yekutieli, D., Letwin, N. E., Elmer, G. I., Lee, N. H., Kafkafi, N. & Benjamini, Y. 2007 Associating quantitative behavioral traits with gene expression in the brain: searching for diamonds in the hay. *Bioinformatics* **23**, 2239–2246. (doi:10.1093/bioinformatics/btm300)

Romano, J. P., Shaikh, A. M. & Wolf, M. 2008 Control of the false discovery rate under dependence using the bootstrap and subsampling (with discussion). *TEST* **17**, 417–442. (doi:10.1007/s11749-008-0126-6)

Storey, J. D. 2002 A direct approach to the false discovery rate. *J. R. Stat. Soc. B* **64**, 479–498. (doi:10.1111/1467-9868.00346)

Storey, J. D., Taylor, J. E. & Siegmund, D. 2003 Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. B* **66**, 187–205. (doi:10.1111/j.1467-9868.2004.00439.X)

Yekutieli, D. 2008 Hierarchical false discovery rate controlling methodology. *J. Am. Stat. Assoc.* **103**, 309–316. (doi:10.1198/016214507000001373)

Yekutieli, D. 2009 Adjusted Bayesian inference for selected parameters. (http://arxiv.org/abs/0801.0499v3)

Zeggini, E. *et al.* 2007 Replication of genome-wide association signals in U.K. samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341. (doi:10.1126/science.1142364)