Discovering Findings That Replicate From a Primary Study of High Dimension to a Follow-Up Study

Marina Bogomolov [a] & Ruth Heller [b]

[a] Technion—Israel Institute of Technology , Technion City , Haifa , 32000 , Israel

[b] Department of Statistics and Operations Research , Tel-Aviv University , Tel-Aviv , Israel

PLEASE SCROLL DOWN FOR ARTICLE

Discovering Findings That Replicate From a Primary Study of High Dimension to a Follow-Up Study

Marina BOGOMOLOV and Ruth HELLER

We consider the problem of identifying whether findings replicate from one study of high dimension to another, when the primary study guides the selection of hypotheses to be examined in the follow-up study as well as when there is no division of roles into the primary and the follow-up study. We show that existing meta-analysis methods are not appropriate for this problem, and suggest novel methods instead. We prove that our multiple testing procedures control for appropriate error rates. The suggested family-wise error rate controlling procedure is valid for arbitrary dependence among the test statistics within each study. A more powerful procedure is suggested for false discovery rate (FDR) control. We prove that this procedure controls the FDR if the test statistics are independent within the primary study, and independent or have positive dependence in the follow-up study. For arbitrary dependence within the primary study, and either arbitrary dependence or positive dependence in the follow-up study, simple conservative modifications of the procedure control the FDR. We demonstrate the usefulness of these procedures via simulations and real data examples. Supplementary materials for this article are available online.

KEY WORDS: False discovery rate; Genome-wide association studies; Meta-analysis; Multiple comparisons; Replicability analysis.

1. INTRODUCTION

In genomics research, it is customary that a primary study is followed by an independent study. Reporting results from the primary study, and then reporting the evidence from the follow-up study that supports these results, gives a sense of the replicability of the results. For example, findings are informally regarded as replicated if the *p*-value for testing a null hypothesis is small in the primary study, and then for the same hypothesis the *p*-value is fairly small in the follow-up study.

Many approaches are available for analyzing two or more studies, where the follow-up studies simply serve to add power. See Hedges and Olkin (1985), Benjamini and Yekutieli (2005), Skol et al. (2006), and Zeggini et al. (2007), among others. In this work, we focus on analyzing two studies, where the follow-up study serves to confirm the findings that were identified in the primary study. A formal statistical approach is proposed for evaluating whether results from a primary study were indeed replicated in a follow-up study.

In observational studies, an association may fail to replicate because the discovered association was not the actual effect of a treatment but rather that of bias (Rosenbaum 2001). However, if the finding is replicated in a different cohort, using different diagnostic or laboratory methods, then the association between effect and outcome may be more convincingly causal. Rosenbaum (2001) gave the example of radiation and leukemia. Suppose higher rates of leukemia are discovered in a primary study among radiologists, and in a follow-up study among survivors at Hiroshima and Nagasaki. Radiation is more convincingly causal if the association discovered was replicated in the follow-up study, since if radiation was not a cause of leukemia, then higher rates of leukemia among radiologists would not lead us to expect higher rates of leukemia among survivors at Hiroshima and Nagasaki. Another example comes from the field

of genomic research. Genome-wide association studies (GWAS) are observational studies, and therefore there is always a danger that bias may explain away the discoveries. Kraft, Zeggini, and Ioannidis (2009) noted that for common variants, the anticipated effects are modest and very similar in magnitude to the subtle biases that may affect genetic association studies—most notably population stratification bias. For this reason, they argued that it is important to see the association in other studies conducted using a similar, but not identical, study base.

It is a common practice that interesting findings in a primary genomic study are investigated in another study, and the interesting results of both studies are reported (Lander and Kruglyak 1995). For example, to discover association between single-nucleotide polymorphisms (SNPs) and hippocampal volume, Bis et al. (2012) tested 2.5×10^6 SNPs in a primary study, and only a handful of SNPs in promising loci in a follow-up study. Bis et al. (2012) forwarded an SNP for replication if the SNP *p*-value in the primary study was below 4×10^{-7}, corresponding to one expected false positive if all SNPs are not associated with hippocampal volume. They viewed the SNP as containing evidence of replication if its *p*-value in the follow-up study was below 0.01, which is the Bonferroni threshold when five hypotheses are simultaneously tested at the 0.05 family-wise error rate (FWER). Their approach selects hypotheses for follow-up based on suggestive evidence (Lander and Kruglyak 1995), and corrects for multiplicity only in the follow-up study when discussing evidence of replicability. Another naive approach is the following: apply a multiple testing procedure within each study separately, and declare as replicated the common findings. This approach will lead to declaring SNPs that were found to be associated with the disease in the primary study as well as in the follow-up study as the discoveries of interest. If there was no danger that a multiple testing procedure produces false positives, then this naive approach would have been appropriate. However, multiple testing procedures have a nonzero probability of producing false positives, unless they have no power. Therefore,

an approach that provides control over false positives in each study separately does not guarantee control over false positives for evaluating whether the results were replicated. Figure 3, left panel, shows that the false discovery rate (FDR) level can be as high as one when naively declaring results as replicated if they were discovered by applying an FDR controlling procedure at the nominal 0.05 level separately in each study. Moreover, reducing the nominal 0.05 level does not resolve the problem, see Remark 3.1.

The article is organized as follows. Section 2 gives the notation and review. Section 3 suggests novel multiple testing procedures for replicability analysis, when the primary study guides the selection of hypotheses to be examined in a follow-up study. Section 4 considers the setting where there is no division of roles into a primary and a follow-up study. In Section 5, we revisit the example of Bis et al. (2012). We also analyze an additional GWAS study, and show additional examples from the GWAS simulator HAPGEN2 (Su, Marchini, and Donnelly 2011). Section 6 describes a simulation study, and Section 7 gives some final remarks.

2. NOTATION, GOAL, AND REVIEW

Consider a family of m elementary null hypotheses $H_1, \ldots, H_m$. These elementary null hypotheses, or a subset thereof, are tested in each of two independent studies. Let h_{ij} be the indicator of whether H_j is false in study i. The pair of indicators (h_{1j}, h_{2j}) identifies four possible settings for each j,

$$(h_{1j}, h_{2j}) = \begin{cases} (0,0) & \text{if } H_j \text{ is true in both studies,} \\ (1,0) & \text{if } H_j \text{ is false in the primary study but} \\ & \quad \text{true in the follow-up study,} \\ (0,1) & \text{if } H_j \text{ is true in the primary study but} \\ & \quad \text{false in the follow-up study,} \\ (1,1) & \text{if } H_j \text{ is false in both studies.} \end{cases}$$

The set of indices $\{1, \ldots, m\}$ of the elementary null hypotheses may be divided into four (unknown) subsets $I_{00} \cup I_{10} \cup I_{01} \cup I_{11} = \{1, \ldots, m\}$, where each index j is in exactly one of the four subsets, defined as follows: $I_{00} = \{j : (h_{1j}, h_{2j}) = (0,0), \quad j \in \{1, \ldots, m\}\}$; $I_{10} = \{j : (h_{1j}, h_{2j}) = (1,0), j \in \{1, \ldots, m\}\}$; $I_{01} = \{j : (h_{1j}, h_{2j}) = (0,1), \quad j \in \{1, \ldots, m\}\}$; $I_{11} = \{j : (h_{1j}, h_{2j}) = (1,1), j \in \{1, \ldots, m\}\}$.

Definition 2.1. The *no replicability null hypothesis* for elementary hypothesis H_j is

$$H_{\text{NR},j} : (h_{1j}, h_{2j}) \in \{(0,0), (0,1), (1,0)\}.$$

By definition, $H_{\text{NR},j}$ is false if and only if the elementary null hypothesis H_j is false in both studies considered. In the family of m composite null hypotheses $H_{\text{NR},1}, \ldots, H_{\text{NR},m}$, the sets of indices of true and false null hypotheses are $I_{00} \cup I_{01} \cup I_{10}$ and I_{11}, respectively. Our goal is to discover as many indices from I_{11} as possible, that is, true positives, while controlling for the number of discoveries from $I_{00} \cup I_{01} \cup I_{10}$, that is, false positives.

Let p_{ij} be the p-value for the jth SNP in study i, for $i = 1, 2$. Since the studies are independent, the p-values are independent across studies. However, the p-values within each study may be

dependent. Inequality $x \geq y$ for vectors x and y is understood componentwise.

Remark 2.1. In a typical meta-analysis (Hedges and Olkin 1985), the goal is to discover as many indices from $I_{01} \cup I_{10} \cup I_{11}$ as possible, while controlling for the number of discoveries from I_{00}. Had we known, and had it been true, that $I_{01} = \emptyset$ and $I_{10} = \emptyset$, then the typical methods for meta-analysis could serve to discover replicable findings. However, it is not known in practice whether I_{01} and I_{10} are empty sets, and they need not be empty when the follow-up study is different, in at least one aspect of design, from the primary study. Therefore, typical meta-analysis methods are not appropriate when the aim is to discover hypotheses with indices in I_{11}, treating all discoveries from I_{01} and I_{10}, in addition to I_{00}, as false discoveries.

2.1 The Partial Conjunction Approach

In Benjamini, Heller, and Yekutieli (2009), the partial conjunction approach (Benjamini and Heller 2008) has been suggested for replicability analysis when $n \geq 2$ studies are available that examine the same problem. When exactly two studies are available, the procedure in Benjamini, Heller, and Yekutieli (2009) amounts to applying the Benjamini–Hochberg (BH) FDR controlling procedure (Benjamini and Hochberg 1995), henceforth referred to as the BH procedure, on the maximum of the two study p-values. However, this procedure may be too conservative, making it practically very difficult to discover false no replicability null hypotheses.

As an example, suppose there is an original GWA study that examines the association of 10^6 SNPs with a phenotype. Now suppose 200 promising SNPs were selected to be examined in a follow-up study. If an SNP has a p-value of $0.025/10^6$ in the first study, and of $0.025/200$ in the second study, then the maximum p-value is $0.025/200$. The BH procedure will, most probably, not reject the no replicability null hypothesis for an SNP with maximum p-value of $0.025/200$, since this maximum p-value is not strong enough evidence when faced with 10^6 hypotheses, out of which most of the hypotheses are true no replicability null hypotheses. The alternative procedures we suggest in Section 3 will view the evidence from this SNP as strong enough for it to be considered a replicated finding.

3. REPLICABILITY ANALYSIS WITH A PRIMARY AND A FOLLOW-UP STUDY

For the family of m no replicability null hypotheses $H_{\text{NR},1}, \ldots, H_{\text{NR},m}$, we consider two relevant error measures: the probability that at least one no replicability null hypothesis is falsely rejected, that is, the FWER, and the expected fraction of false rejections out of all rejections of no replicability null hypotheses, that is, the FDR.

Procedure 3.1. The two-stage FWER controlling procedure for testing the family of no replicability null hypotheses with parameters (α_1, α), where $0 < \alpha_1 < \alpha < 1$:

1. Let $\mathcal{R}_1$ be the set of indices of elementary hypotheses that are selected for testing in a follow-up study based on the data from the primary study.

2. Apply an FWER controlling procedure at level α_1, using the data from the primary study only, on the family of null hypotheses $H_1, \ldots, H_m$, and let $\mathcal{R}_p \subseteq \{1, \ldots, m\}$ be the set of indices of rejected hypotheses. Apply an FWER controlling procedure at level $\alpha - \alpha_1$, using the data from the follow-up study only, on the family of selected null hypotheses $\{H_j : j \in \mathcal{R}_1\}$, and let $\mathcal{R}_f \subseteq \mathcal{R}_1$ be the set of indices of rejected hypotheses. Then the set of indices of rejected no replicability null hypotheses is $\mathcal{R}_f \cap \mathcal{R}_p$.

Theorem 3.1. For two independent studies, Procedure 3.1 controls the FWER at level α for the family of no replicability null hypotheses $H_{NR,1}, \ldots, H_{NR,m}$.

Proof. Let $V_p = \sum_{j \in \mathcal{R}_p}(1 - h_{1j})$ and $V_f = \sum_{j \in \mathcal{R}_f}(1 - h_{2j})$ be the number of true elementary null hypotheses rejected, respectively, in the primary study and in the follow-up study. Then

$$\begin{aligned} \text{FWER} &\leq E(\mathbf{I}[V_p + V_f > 0]) \leq E(\mathbf{I}[V_p > 0]) \\ &+ E(E(\mathbf{I}[V_f > 0]|(p_{11}, \ldots, p_{1m}))) \leq \alpha_1 + \alpha - \alpha_1 = \alpha, \end{aligned}$$

where the last inequality follows from the fact that V_f is independent of the data from the primary study, and that in both studies an FWER controlling procedure is applied. $\square$

Using Bonferroni in Procedure 3.1 amounts to rejecting $H_{NR,j}$ if $(p_{1j}, p_{2j}) \leq (\alpha_1/m, (\alpha - \alpha_1)/|\mathcal{R}_1|)$, for $j \in \mathcal{R}_1$. Alternatively, the results can be reported in terms of Bonferroni-replicability adjusted p-values $p_j^{\text{Bonf}-\text{REPadj}} = \max(mp_{1j}/c, |\mathcal{R}_1|p_{2j}/(1 - c))$, where $c = \alpha_1/\alpha$. Procedure 3.1 using Bonferroni is equivalent to rejecting all hypotheses with Bonferroni-replicability adjusted p-values at most α.

The selection rule affects the power of Procedure 3.1. A natural choice for a selection rule is the set of rejected hypotheses by the FWER controlling procedure at level α_1 on the primary study p-values, since the set of indices of rejected no replicability null hypotheses is a subset of this set. A rule that selects by the FWER controlling procedure at level α is not as good, since any additional hypotheses selected will not be rejected but will result in a more severe multiple testing problem for the follow-up study. The choice of α_1 also affects the power of Procedure 3.1. The optimal choice depends on the nonnull distribution of the p-values, as well as on $|I_{00}|, |I_{10}|, |I_{01}|, |I_{11}|$, and the FWER controlling procedure used. Therefore, guidelines for choosing α_1 are application specific. In simulations (supplementary material) the power function was quite flat as long as α_1/α was not too close to zero or one.

In many modern applications, controlling the FWER is unnecessary and results in overly conservative inferences. In genomics research, it is often enough to guarantee FDR control, see Storey and Tibshirani (2003) and Reiner, Yekutieli, and Benjamini (2003), among others.

Procedure 3.2. The two-stage FDR controlling procedure for testing a family of no replicability null hypotheses with parameters (q_1, q), where $0 < q_1 < q < 1$:

1. Let $\mathcal{R}_1$ be the set of indices of elementary hypotheses that are selected for testing in a follow-up study based on the data from the primary study. Let $R_1 = |\mathcal{R}_1|$ be the cardinality of this set.

2. Let

$$\begin{aligned} R_2 &\triangleq \max \left\{ r : \sum_{j \in \mathcal{R}_1} \mathbf{I}\left[(p_{1j}, p_{2j}) \right. \right. \\ &\left. \left. \leq \left(\frac{rq_1}{m}, \frac{r(q - q_1)}{R_1} \right) \right] = r \right\}. \end{aligned}$$

Then the set of indices of rejected no replicability null hypotheses is

$$\mathcal{R}_2 = \left\{ j : (p_{1j}, p_{2j}) \leq \left(\frac{R_2 q_1}{m}, \frac{R_2(q - q_1)}{R_1} \right), j \in \mathcal{R}_1 \right\}.$$

The results of Procedure 3.2 can be reported in terms of FDR-replicability adjusted p-values. Let $c = q_1/q$,

$$Z_j = \max \left(\frac{m p_{1j}}{c}, \frac{R_1 p_{2j}}{1 - c} \right), j \in \mathcal{R}_1, \qquad (3.1)$$

and let $Z_{(1)} \leq \cdots \leq Z_{(R_1)}$ be the sorted Z-values. Then the ith largest FDR-replicability adjusted p-value is

$$p_{(i)}^{\text{REPadj}} = \min_{j \geq i} \frac{Z_{(j)}}{j}. \qquad (3.2)$$

Procedure 3.2 with parameters $(q_1, q) = (cq, q)$ is equivalent to rejecting all no replicability null hypotheses with FDR-replicability adjusted p-values at most q.

Definition 3.1. A *stable selection rule* for Step 1 of Procedure 3.2 satisfies the following condition: for any $j \in \mathcal{R}_1$, fixing all the p-values except for p_{1j} and changing p_{1j} so that H_{1j} is still selected will not change the set $\mathcal{R}_1$.

It is easy to see that this condition is satisfied if $\mathcal{R}_1$ contains the smallest fixed number of p-values, all hypotheses with p-value below a given threshold, or if $\mathcal{R}_1$ contains the rejected indices from a BH procedure on the p-values from the primary study. Adaptive FDR procedures on the p-values from the primary study, for example, Benjamini and Hochberg (2000), Storey, Taylor, and Siegmund (2004), Benjamini, Krieger, and Yekutieli (2006), and Blanchard and Roquain (2009), are nonstable selection rules.

Theorem 3.2. If all the p-values are jointly independent and the selection rule in step 1 of Procedure 3.2 is a stable selection rule, then Procedure 3.2 controls the FDR at level q for the family of no replicability null hypotheses $H_{NR,1}, \ldots, H_{NR,m}$.

See Appendix A for the proof.

The selection rule and the choice of q_1 affect the power of Procedure 3.2. A natural choice for a selection rule is the set of rejected hypotheses by the BH procedure at level q_1 on the primary study p-values, since the set of indices of rejected no replicability null hypotheses is a subset of this set. A rule that selects by the BH procedure at level q is not as good as the rule at level q_1, since any additional hypotheses selected will not be rejected, but will result in more severe thresholds on the follow-up study p-values. In Figure 4, we showed in a simulated example that the BH procedure at level q_1 was very close to selecting the optimal number of hypotheses for follow-up. We recommend using it when there are no additional constraints that require choosing only a small number of hypotheses for follow-up. The optimal choice of q_1 depends on $|I_{00}|, |I_{01}|, |I_{10}|, |I_{11}|$, and the

nonnull distribution of the p-values, and therefore guidelines for choosing q_1 are application specific. In simulated GWAS in Section 5, the choice of q_1 had little effect on the average number of discoveries.

Theorem 3.2 assumes independence of the p-values within each study as well as across the studies. However, the assumption of independence among the p-values within each study may not be realistic in many applications. Particularly, in GWAS there is dependency across the SNPs, therefore the p-values within each study may be dependent. Benjamini and Yekutieli (2001) proved that the BH procedure controls the FDR when the p-values have a special dependency called PRDS.

Definition 3.2. (Benjamini and Yekutieli 2001). The set of p-values $P_1, \ldots, P_M$ has property PRDS if for any increasing set D, and for each true null hypothesis i, $\Pr((P_1, \ldots, P_M) \in D | P_i = p)$ is nondecreasing in p.

If the p-values are independent in the primary study, yet have property PRDS in the follow-up study, Theorem S3.1 in the supplementary material shows that the result in Theorem 3.2 holds. For arbitrary dependence among the p-values in the primary study, a modification of the cut-off level of Procedure 3.2 will guarantee that the FDR is controlled at the nominal level. The most severe modification, which will guarantee FDR control for any stable selection rule, is to apply Procedure 3.2 with the modification in item 1 of Theorem 3.3. However, in item 2 of Theorem 3.3 we show that the modification factor may be smaller than $\sum_{i=1}^{m} 1/i \approx \log m$ if the selected hypotheses for follow-up are a subset of the hypotheses with primary study p-values below a fixed cut-off t. For example, in GWAS it is common to select hypotheses with primary study p-values below $1/m$, where m is the number of hypotheses in the primary study (Lander and Kruglyak 1995). If $t \geq \frac{q_1}{1 + \sum_{i=1}^{m-1} \frac{1}{i}}$, then the modification in item 1 of Theorem 3.3 cannot be improved. However, if $t < \frac{q_1}{1 + \sum_{i=1}^{m-1} \frac{1}{i}}$, then the modification in item 2 of Theorem 3.3 is less conservative than the modification in item 1. For typical values of q_1 (e.g., $q_1 \in [0.005, 0.045]$) and large m, the threshold t will often be below $\frac{q_1}{1 + \sum_{i=1}^{m-1} \frac{1}{i}}$, and therefore item 2 may be useful in applications. Note, moreover, that if $t \leq q_1/m$, then item 2 of Theorem 3.3 states that no modification is required, so for a stable selection rule that selects a subset of the set of hypotheses with primary study p-values below t, where $t \leq q_1/m$, Procedure 3.2 is valid for any form of dependency among the p-values in the primary study.

Theorem 3.3. Assume that the follow-up study p-values have property PRDS, and are independent of the p-values in the primary study. Then Procedure 3.2 controls the FDR at level q for the family of no replicability null hypotheses $H_{NR,1}, \ldots, H_{NR,m}$ if the selection rule used in Step 1 of Procedure 3.2 is a stable selection rule, and the expressions in Step 2 of Procedure 3.2 are modified as follows:

1. In the terms rq_1/m and R_2q_1/m only, q_1 is replaced by $\widetilde{q}_1 = q_1/(\sum_{i=1}^{m} 1/i)$.

2. In the terms rq_1/m and R_2q_1/m only, q_1 is replaced by $\widetilde{q}_1$, where

$$\widetilde{q}_1 = \max\left\{x : \ x\left(1 + \sum_{i=1}^{\lceil tm/x - 1\rceil} 1/i\right) = q_1\right\},$$

if only hypotheses with primary study p-values at most a fixed threshold t are considered for follow-up, that is, $\mathcal{R}_1 \subseteq \{j \in \{1, \ldots, m\} : \ P_{1j} \leq t\}$, where $t < \frac{q_1}{1 + \sum_{i=1}^{m-1} \frac{1}{i}}$.

See supplementary material for the proof, as well as for additional results under dependency. Specifically, Theorem S3.2 in the supplementary material shows that in the more general setting of arbitrary dependence among the follow-up study p-values, it is also necessary to replace $(q - q_1)$ with $(q - q_1)/(\sum_{i=1}^{R_1} 1/i)$ in the terms $r(q - q_1)/R_1$ and $R_2(q - q_1)/R_1$ in expression 2 of Procedure 3.2. These results are similar to the result in Benjamini and Yekutieli (2001) for the BH procedure in their Theorem 1.3.

Remark 3.1. Benjamini and Yekutieli (2005) proved in their Proposition 3 that the procedure that applies the BH procedure at level q_1 on the primary study p-values, and the BH procedure at level $q - q_1$ on the follow-up study p-values controls the FDR at level $q_1(q - q_1) < q$ on the family of global null hypotheses, $H_{G1}, \ldots, H_{Gm}$, where $H_{Gj} : (h_{1j}, h_{2j}) = (0, 0)$. However, on the family of no replicability null hypotheses, $H_{NR,1}, \ldots, H_{NR,m}$, the FDR of this procedure may be higher than the nominal level q. The key difference between Procedure 3.2 and such a two-stage procedure is the requirement that the two p-values from a selected hypothesis have to simultaneously be smaller than two thresholds. In an extreme scenario where all hypotheses are from I_{10} or I_{01}, and the p-values from false null hypotheses are zero, the two-stage procedure may have an FDR of one, as follows. The BH procedure on the primary study p-values will reject all hypotheses from I_{10} but also few from I_{01} (when $|I_{01}|$ and $|I_{10}|$ are large enough), and the hypotheses from I_{01} will be rejected by the BH procedure on the follow-up study p-values, resulting in an FDR of one. However, Procedure 3.2 will have an FDR level below q. To see this, note that to reject a no replicability null hypothesis by Procedure 3.2, the p-value of the Simes (1986) test for the intersection of the elementary hypotheses indexed by I_{01}, using the data from the primary study, has to be below q_1, or the Simes test p-value for the intersection of elementary hypotheses indexed by $I_{10} \cap \mathcal{R}_1$, using the data from the follow-up study, has to be below $q - q_1$. Therefore, the probability of rejecting at least one no replicability null hypothesis, which coincides with the FDR since all no replicability null hypotheses are true, is at most q. See Figure 3, right panel, for a more realistic simulated example.

4. REPLICABILITY ANALYSIS WITH NO DIVISION INTO PRIMARY AND FOLLOW-UP STUDIES

Consider now a situation where both studies are available before the analysis. If some of the elementary hypotheses are examined in only one of the studies, then these hypotheses are not considered for replicability analysis. In this setting, there is no primary study and follow-up study. We propose the following generalization of Procedure 3.2, which can be tuned to treat the

two studies symmetrically. Without loss of generality, we label the studies as study one and study two.

Procedure 4.1. The generalized two-stage procedure for testing a family of no replicability null hypotheses with parameters (w_1, q_1, q), where $0 \leq w_1 \leq 1$ and $0 < q_1 < q < 1$:

1. Apply Procedure 3.2 with parameters $(w_1 q_1, w_1 q)$ with study one as the primary study and study two as the follow-up study. Denote the set of indices of rejected no replicability null hypotheses by $\mathcal{R}_{12, w_1 q}$.
2. Reverse the roles of study one and study two. Apply Procedure 3.2 with parameters $((1 - w_1)q_1, (1 - w_1)q)$. Denote the set of indices of rejected no replicability null hypotheses by $\mathcal{R}_{21, (1-w_1)q}$.
3. The set of indices of rejected no replicability null hypotheses is $\mathcal{R}_{12, w_1 q} \cup \mathcal{R}_{21, (1-w_1)q}$.

Theorem 4.1. Procedure 4.1 controls the FDR at level q for the family of no replicability null hypotheses $H_{\mathrm{NR},1}, \ldots, H_{\mathrm{NR},m}$ if all p-values are jointly independent and the selection rule in Step 1 of Procedure 3.2 is a stable selection rule.

See Appendix C for the proof.

Choosing $w_1 = 1$ results in Procedure 3.2, where study one has the role of the primary study and study two has the role of the follow-up study. Similarly, choosing $w_1 = 0$ results in Procedure 3.2 with the roles of study one and study two reversed. The choice $0 < w_1 < 1$ reflects the similarity of Procedure 4.1 to Procedure 3.2 in the following way: when w_1 is close to one (zero), Procedure 4.1 gives similar results to Procedure 3.2 with study one (two) as the primary study. The choice $w_1 = 0.5$ results in a variant of Procedure 3.2 that is symmetric with respect to both studies.

5. GWAS EXAMPLES

In this section, we demonstrate the suggested methods on two real data examples and on a GWAS simulation. A replicability analysis with FWER control is carried out for the first example that has only five hypotheses in the follow-up study. A replicability analysis with FDR control is carried out for the second example that has 126 hypotheses in the follow-up study. Finally, to examine the robustness of procedure 3.2 for GWAS type dependency, examples were simulated that retained the dependencies in the data that occur in GWAS.

Example 1. We reproduce in Table 1, columns 1–4, a subset of the columns of Table 1 of results of Bis et al. (2012). We added in columns 5–7, the Bonferroni-replicability adjusted p-values

for $c = q_1/q \in \{0.2, 0.5, 0.8\}$. Procedure 3.1 with parameters $(q_1, q) = (0.025, 0.05)$ or $(q_1, q) = (0.04, 0.05)$ identified the SNP near MSRB3 as having replicated association with the phenotype. The choice of c should be made prior to analysis, and the choice $c = 0.8$ may be preferred over $c \leq 0.5$ when it is believed that the power to detect an association in the primary study using a threshold of order $1/(2.5 \times 10^6)$ is smaller than the power to detect an association in the follow-up study using a threshold of order $1/5$.

Example 2. To discover associations between SNPs and Crohn's disease (CD), Barrett et al. (2008) examined 635,547 SNPs on 3230 cases and 4829 controls of European descent, collected in three separate studies: NIDDK4, WTCCC5, and a Belgian-French study. The primary study p-values in this example are the meta-analysis p-values from the combined data from the three studies. Only hypotheses with primary study p-values below 5×10^{-5} were considered for follow-up. Although 526 SNPs met the selection criterion, only a subset of 126 SNPs were followed up. These 126 p-values were the smallest two p-values in 63 distinct regions, so the selection rule is a stable selection rule. Procedure 3.2 with $(q_1, q) = (0.04, 0.01)$ identified 36 SNPs. In Appendix D, Table D.1 shows the p-values from the primary and follow-up studies, as well as the FDR-replicability adjusted p-values for the choice $c = 0.8$, for these 36 replicability discoveries. Since the p-values are not independent within each study, a more conservative analysis approach is to modify the cut-offs as suggested by Theorem 3.3. Assuming PRDS type dependency in the follow-up study, item 1 of Theorem 3.3 suggests using $\tilde{q}_1 = 0.04/(\sum_{i=1}^{635,547} 1/i) = 0.0029$ for the primary study cut-offs, while the follow-up study cut-offs remain unchanged. The modified procedure identified 21 SNPs. Column 7 of Table D.1 shows the FDR-replicability adjusted p-values for the choice $c = 0.8$, where the adjustment is made as described in expressions (3.1) and (3.2), with p_{1j} replaced by $\tilde{p}_{1j} = (\sum_{i=1}^{635,547} 1/i)p_{1j} = 13.94 \times p_{1j}$, $j = 1, \ldots, 635,547$. Since the SNPs considered for follow-up were only SNPs with primary study p-values below 5×10^{-5}, one could use a less conservative procedure suggested in item 2 of Theorem 3.3, with $\tilde{q}_1 = 0.0038$, where 0.0038 is the solution to $0.04 = x(\sum_{i=1}^{\lceil 635,547 \times 5 \times 10^{-5}/x - 1 \rceil} 1/i + 1)$. This procedure resulted in 23 replicability discoveries. The latter procedure is the recommended procedure, if the investigator is not willing to assume that Procedure 3.2 is robust to deviations from independence within the primary study. However, simulations in the next example suggest that for the type of dependencies that occur in GWAS, Procedure 3.2 may actually be conservative.

Table 1. The p-values of SNPs from the primary and follow-up studies, from Table 1 of Bis et al. (2012) (columns 3 and 4), and the Bonferroni-replicability adjusted p-values for various choices of $c = q_1/q$ (columns 5–7)

| Locus | Gene | Primary study | Follow-up study | Bonferroni-replicability adjusted p-values | | |
				$c = 0.2$	$c = 0.5$	$c = 0.8$
2q24	DPP4	5.2×10^{-8}	0.7	1.0000	1.0000	1.0000
9q33	ASTN2	1.0×10^{-7}	0.2	1.0000	0.5000	0.3125
12q14	MSRB3	5.5×10^{-9}	0.002	0.06875	0.0275	0.0172
	WIF1	2.2×10^{-8}	0.0007	0.2750	0.1100	0.0688
12q24	HRK	4.8×10^{-8}	5.8×10^{-5}	0.6000	0.2400	0.1500

Table 2. For 4500 cases and 4500 referents in both studies, the average number of associated and disease SNPs discovered (SE), and the average FDP (SE), for different procedures. The selection rule for Procedure 4.1 was the BH procedure at level $w_1 q_1$ when the CEU study was the primary study, and at level $(1 - w_1)q_1$ when the CHD study was the primary study

| | No. of replicated findings | | |
Procedure	Associated SNPs(SE)	Disease SNPs(SE)	FDP (SE)
BH on maximum p-values	29.182 (3.205)	7.364 (0.432)	0.000 (0.000)
4.1 with $w_1 = 0.5, q_1 = 0.025, q = 0.05$	77.727 (6.378)	11.455 (0.366)	0.011 (0.005)
4.1 with $w_1 = 1, q_1 = 0.01, q = 0.05$	74.091 (6.748)	10.364 (0.310)	0.012 (0.006)
4.1 with $w_1 = 1, q_1 = 0.025, q = 0.05$	76.091 (6.221)	10.727 (0.359)	0.012 (0.005)
4.1 with $w_1 = 1, q_1 = 0.04, q = 0.05$	69.545 (5.745)	10.818 (0.352)	0.009 (0.005)
4.1 with $w_1 = 0, q_1 = 0.01, q = 0.05$	35.545 (4.575)	7.364 (0.607)	0.008 (0.008)
4.1 with $w_1 = 0, q_1 = 0.025, q = 0.05$	41.455 (5.294)	8.273 (0.469)	0.007 (0.007)
4.1 with $w_1 = 0, q_1 = 0.04, q = 0.05$	42.273 (4.158)	8.545 (0.312)	0.000 (0.000)

We come back to the issue of robustness of Procedure 3.2 in Section 7.

GWAS Simulation Example. We simulated two GWAS from the simulator HAPGEN2 (Su, Marchini, and Donnelly 2011). The two studies were generated from two samples of the HapMap project (The International HapMap Consortium 2003), a sample of 165 Utah residents with Northern and Western European ancestry (abbreviation: CEU), and a sample of 109 Chinese in Metropolitan Denver, Colorado (abbreviation: CHD). In the CEU and CHD populations, respectively, 34 and 38 SNPs were set as disease SNPs with an increased multiplicative relative risk of 1.2, and 18 of the disease SNPs were common to both populations. Each study contained 4500 cases and 4500 referents. The linkage disequilibrium (LD) across SNPs, as measured for the samples in the HapMap project, was retained. Due to LD, the number of SNPs associated with the phenotype in each study was larger than the number of disease SNPs. To identify the SNPs in each study that are truly associated with the phenotype, the simulation of 4500 cases and 4500 controls from the population was repeated 11 times, and 11 p-values were produced per SNP. SNPs with Fisher's combined p-value (Loughin 2004) below the Bonferroni threshold were considered to be truly associated with the disease. Our ground truth included 1355 and 1010 SNPs associated with the disease in the CEU and in the CHD population, respectively, out of which 274 SNPs were associated with the disease in both populations.

As a standard preprocessing step, we removed SNPs with minor allele frequency below 0.05, and thus the number of SNPs in the analysis was reduced from 1,387,466 to 887,362, on average, for the 11 pairs of studies. Our selection rule for Procedure 4.1 with parameters (w_1, q_1, q) was the BH procedure at level $w_1 q_1$ when the primary study was the CEU study, and at level $(1 - w_1)q_1$ when the primary study was the CHD study, since the potential set of SNPs to be discovered as having replicated associations is at most the set of SNPs that are discovered by the BH procedure (as discussed in Section 3). Table 2 presents the average number of replicated findings, as well as the average false discovery proportion (FDP) for the methods compared. The standard error (SE) is presented in parentheses. From rows 1 and 2, we see that if there is no division into primary and follow-up studies, then the symmetric Procedure 4.1 discovers more SNPs with replicated associations than the BH procedure

on maximum p-values, while maintaining a low FDP. From rows 3–5 and 6–8, we see that the choice of which study was the primary study had a large effect on the average number of discoveries, and the choice of q_1 mattered little.

From the last column in Table 2, we see that the average FDP was far below 0.05, suggesting that the procedures are conservative. This conservatism can be alleviated if the following oracle information were known: the fraction of SNPs with no association with the phenotype in both studies, f_{00}, and with association with the phenotype only in the follow-up study, f_{01}. Then it was possible to perform Procedure 4.1 at level $(w_1, q', 2q')$, where q' is the solution to $f_{00}(q')^2 + (f_{01} + 1)q' = q$ for $w_1 \in \{0, 1\}$, and the solution to $f_{00}(0.5q')^2 + (f_{01} + 1)0.5q' = 0.5q$ for $w_1 = 0.5$, with the same guarantee of FDR control at level q, as follows from Appendix B. Specifically, in our simulation $f_{00} = 0.9990$, $f_{01} = 0.00036$ on average, after preprocessing. For FDR control at level $q = 0.05$, on average $q' = 0.048$ for $w_1 = 0, 1$ and $q' = 0.049$ for $w_1 = 0.5$. Table 3 shows the average FDP and average number of rejections for Procedure 4.1 with and without the oracle. Although the average FDP is higher with the oracle, it is still below the nominal 0.05 level for two main reasons. First, our simulation preserves the LD pattern of the SNPs, and thus the p-values within each study are not independent. Second, the upper bound of $f_{00}(q')^2 + (f_{01} + 1)q'$ is not a tight upper bound for the actual FDR level. A tighter oracle upper bound requires knowing the expectation of $|\mathcal{R}_1 \cap I_{10}|/|\mathcal{R}_1|$, and this bound is tight if the nonnull effect sizes in $I_{10} \cup I_{01}$ are extremely large.

Table 3. The average FDP and average number of rejections for Procedure 4.1 with and without the oracle, for FDR control at level 0.05

	FDP		No. of replicated findings	
	Oracle	$(q_1, q) =$ (0.025, 0.05)	Oracle	$(q_1, q) =$ (0.025, 0.05)
$w_1 = 0.5$	0.023	0.011	90	78
$w_1 = 1$	0.023	0.012	85	76
$w_1 = 0$	0.029	0.007	50	41

For the two studies from the CEU and CHD populations, a meta-analysis was performed by first combining the SNP p-values using Fisher's combining method, and then applying the BH procedure at level 0.05 on the combined p-values. The average number of SNPs associated with the disease in at least one study was 393, while less than 80 SNPs were discovered to have replicated associations (Table 2). The two main reasons for discovering more SNPs in a typical meta-analysis are as follows. First, the simulation setting contained five times more associated SNPs than SNPs with replicated associations. Second, for an SNP with a replicated association, the power to detect that the association is replicated is lower than the power to detect that there is an association in at least one study. The discovered SNPs with replicated associations were a subset of the discovered associated SNPs, but their meta-analysis p-values were not ranked smallest among all meta-analysis p-values (not shown). Importantly, the discoveries from the meta-analysis could not serve as evidence toward replicability, since while the average fraction of SNPs with no association in both studies among the meta-analysis discoveries was 0.06, the average fraction of SNPs with no replicated association among the meta-analysis discoveries was 0.78.

6. A SIMULATION STUDY

The goal of the simulations was threefold. First, to investigate the effect of the choice of q_1 and w_1 on the power of Procedures 3.2 and 4.1. Second, to compare these procedures to the alternative of applying BH on the maximum p-values, that is, the partial conjunction approach when exactly two studies are analyzed. Third, to investigate the effect of the selection rule on the power of the procedures.

The procedures compared were (1) the BH procedure at level 0.05 on maximum p-values; (2) Procedure 4.1 with $w_1 \in \{0, 0.5, 1\}$, $c = q_1/q \in \{0.1, 0.2, \dots, 0.9\}$, and $q = 0.05$;

and (3) the naive (BH-i, BH-j) procedure, $i, j \in \{1, 2\}$, $i \neq j$, which applies the BH procedure at level 0.05 on the p-values of study i, and separately on the p-values of study j for the hypotheses that were rejected in study i, and declares hypotheses rejected in both studies as false no replicability null hypotheses; (4) the oracle Procedure 3.2 with parameters $(q_1, q) = (q', 2q')$, where q' was the solution to $\frac{|I_{00}|}{m}(q')^2 + (\frac{|I_{01}|}{m} + 1)q' = 0.05$. This oracle procedure controls the FDR at level 0.05, see Appendix B for a proof.

The p-values were generated independently as follows. For H_j, $j = 1, \dots, m$, $P_{1j} = 1 - \Phi(\frac{X_{1j}}{\sigma_1})$, and $P_{2j} = 1 - \Phi(\frac{X_{2j}}{\sigma_2})$, where $X_{1j} \sim N(\mu_{1j}, \sigma_1^2)$ and $X_{2j} \sim N(\mu_{2j}, \sigma_2^2)$. We let $\mu_{ij} = 0 \cdot (1 - h_{ij}) + \mu_i \cdot h_{ij}$, where $i \in \{1, 2\}$, and $\mu_i \in \{0.5, 1, \dots, 5\}$. We set $m = 1000$, and $f_{ij} = |I_{ij}|/m$ for $i, j \in \{0, 1\}$ as follows: $f_{00} = 0.9$, $f_{11} = 0.1$; $f_{00} = 0.9$, $f_{01} = f_{10} = 0.025$, $f_{11} = 0.05$; $f_{01} = f_{10} = 0.5$; $f_{00} = 0.8$, $f_{01} = f_{10} = 0.1$. The standard deviations σ_1 and σ_2 were either fixed values $\sigma_i \in \{0.3, 1\}$, $i \in \{1, 2\}$, or reflected the fraction of sample size allocated to the first study: $\sigma_1 = \sigma/\sqrt{\zeta N}$, $\sigma_2 = \sigma/\sqrt{(1 - \zeta)N}$, $\sigma = 10$, $\zeta \in \{0.1, 0.2, \dots, 0.9\}$, $N = 1000$.

The simulation results were based on 1000 repetitions. The FDR was estimated by averaging the FDP. The average power was estimated by the average number of rejected false no replicability null hypotheses, divided by mf_{11}.

6.1 Simulation Results

As expected from our theoretical results, in all the settings considered the estimated FDR was below 0.05 for all procedures but the naive (BH-i, BH-j) procedure. The SE of the estimated FDR and power were of the order of 10^{-3} for all procedures under all configurations considered.

Figure 1 compares the power of the BH procedure on maximum p-values, (1), and Procedure 4.1 with $w_1 \in \{0, 0.5, 1\}$, $q_1 \in \{0.01, 0.025, 0.04\}$, (2), in a configuration with parameters

| (a) q_1=0.01 | (b) q_1=0.025 | (c) q_1=0.04 |
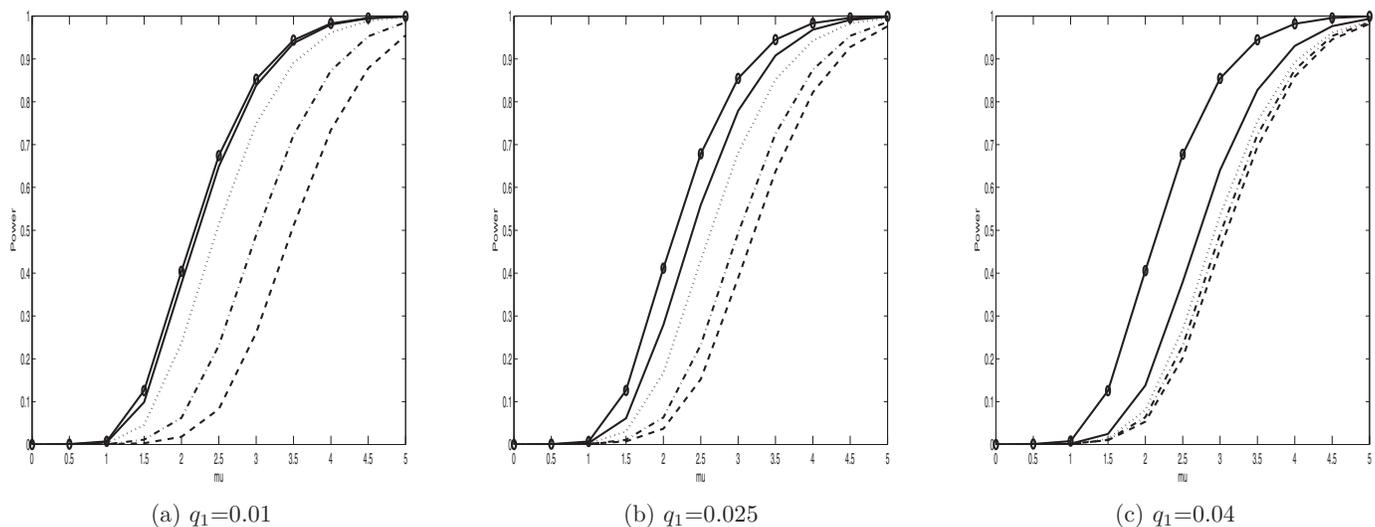
Figure 1. Power as a function of $\mu = \mu_1 = \mu_2$, for q_1 of (a) 0.01, (b) 0.025, and (c) 0.04, using the following procedures: the oracle Procedure 3.2 (solid with circles); the BH procedure at level 0.05 applied on maximum p-values (dash-dotted); Procedure 4.1 at level 0.05 with $w_1 = 0$ (dashed), $w_1 = 0.5$ (dotted), and $w_1 = 1$ (solid), where the selection rule in Steps 1 and 2 is the BH procedure at levels $w_1 q_1$ and $(1 - w_1)q_1$, respectively. The remaining parameters were $f_{00} = 0.9$, $f_{01} = 0.025$, $f_{10} = 0.025$, $f_{11} = 0.05$, $\mu_1 = \mu_2 = \mu$, $\sigma_1 = 0.3$, and $\sigma_2 = 1$.

Table 4. The power of Procedure 3.2 with parameters $(0.05c, 0.05)$ and the BH selection rule at level $0.05c$, for different values of $\mu = \mu_1 = \mu_2$, with $\sigma_1 = \sigma_2 = 0.5$, $f_{00} = 0.9$, $f_{01} = f_{10} = 0.025$, $f_{11} = 0.05$. The optimal value of c is in bold

	$c = q_1/0.05$								
μ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1.5	0.143	0.195	0.224	0.245	0.257	**0.258**	0.248	0.226	0.181
2.0	0.646	0.718	0.755	0.778	0.794	0.803	**0.805**	0.800	0.769
2.5	0.934	0.955	0.965	0.971	0.975	0.977	**0.978**	**0.978**	0.974

$\sigma_1 = 0.3, \sigma_2 = 1, f_{00} = 0.9, \quad f_{01} = f_{10} = 0.025, f_{11} = 0.05$. The oracle Procedure 3.2, where the primary study is study one with $\sigma_1 = 0.3$, is also examined. For each procedure, the estimated power and FDR is shown as a function of the common expectation under the alternative, $\mu = \mu_1 = \mu_2$. Procedure 4.1 with $w_1 = 1$ is more powerful than with $w_1 = 0.5$ or $w_1 = 0$, while the choice $w_1 = 0$ is the worst in terms of power of Procedure 4.1. Moreover, Procedure 4.1 with $w_1 \in \{0.5, 1\}$ is more powerful than the BH procedure on maximum p-values. These findings were consistent across all configurations of $f_{00}, f_{10}, f_{01}, f_{11}$ examined, when $\sigma_1 = 0.3$ and $\sigma_2 = 1$. Since the oracle Procedure 3.2 and the BH procedure on maximum p-values do not depend on q_1, their power curves are the same in Figure 1(a)–1(c). We see that Procedure 4.1 with $w_1 = 1$ is a close second to the oracle when q_1 is 0.01 but is farther from oracle as q_1 increases. Similarly, the power of Procedure 4.1 with $w_1 = 0.5$ decreases as q_1 increases. However, Procedure 4.1 with $w_1 = 0$ has largest power for $q_1 = 0.04$, and the least power for $q_1 = 0.01$. These results are reasonable since the p-values of study one tend to be much smaller than the p-values of study two

when the no replicability null hypotheses are false. In Table 4, we see that if the p-value distribution of false no replicability null hypotheses is the same across studies, then the optimal choice of q_1 is $q_1 > q/2$. For example, when $\mu = \mu_1 = \mu_2 = 2$ (row 2), the power is 0.65 with $q_1 = 0.005$, 0.77 with $q_1 = 0.045$, and the maximum power is 0.81 with $q_1 = 0.035$.

Figure 2 compares the procedures (1) and (2) for the same configuration of f_{ij}, but for fixed $\mu = \mu_1 = \mu_2$ and varying sample size of the two studies. The varying power is described by the fraction ζ of sample allocated to the first study. For the symmetric procedures, we see that for $\zeta = 0.1$ the power is the lowest, and it increases to reach its maximum for equal allocation $\zeta = 0.5$. Procedure 4.1 with $w_1 = 0.5$ dominates the BH procedure on the maximum two study p-values. For Procedure 4.1 with $w_1 = 1$, the maximum is reached for $\zeta > 0.5$. It is the most powerful of the three procedures examined for $\zeta > 0.6$.

In Figure 3, we consider the FDR level of Procedure 4.1 with $w_1 \in \{0, 0.5, 1\}$, as well as of the naive procedure in the null setting, where all no replicability null hypotheses are true (i.e., $f_{11} = 0$). The estimated FDR of (BH-i, BH-j) procedure exceeds 0.05 in the settings where $f_{10} = f_{01} = 0.5$ and $f_{00} = 0.8$, $f_{10} = f_{01} = 0.1$. In these settings, the estimated FDR of both (BH-1, BH-2) and (BH-2, BH-1) procedures are increasing functions of $\mu = \mu_1 = \mu_2$, reaching one in the setting where $f_{10} = f_{01} = 0.5$ (left), and 0.4 in the setting where $f_{00} = 0.8$, $f_{10} = f_{01} = 0.1$ (right). Clearly, procedure (BH-i, BH-j) is not valid since it may be far too liberal in terms of FDR level.

Finally, we examined how the selection rule affects the power. In Figure 4, we show the power as a function of μ_1 for Procedure 4.1 with parameters $w_1 = 0.5$, $q_1 = 0.025$, $q = 0.05$, for the following selection rules: BH at level 0.0125; the rule that selects the hypotheses with k smallest primary

(a) $\mu = 2$ (b) $\mu = 3$
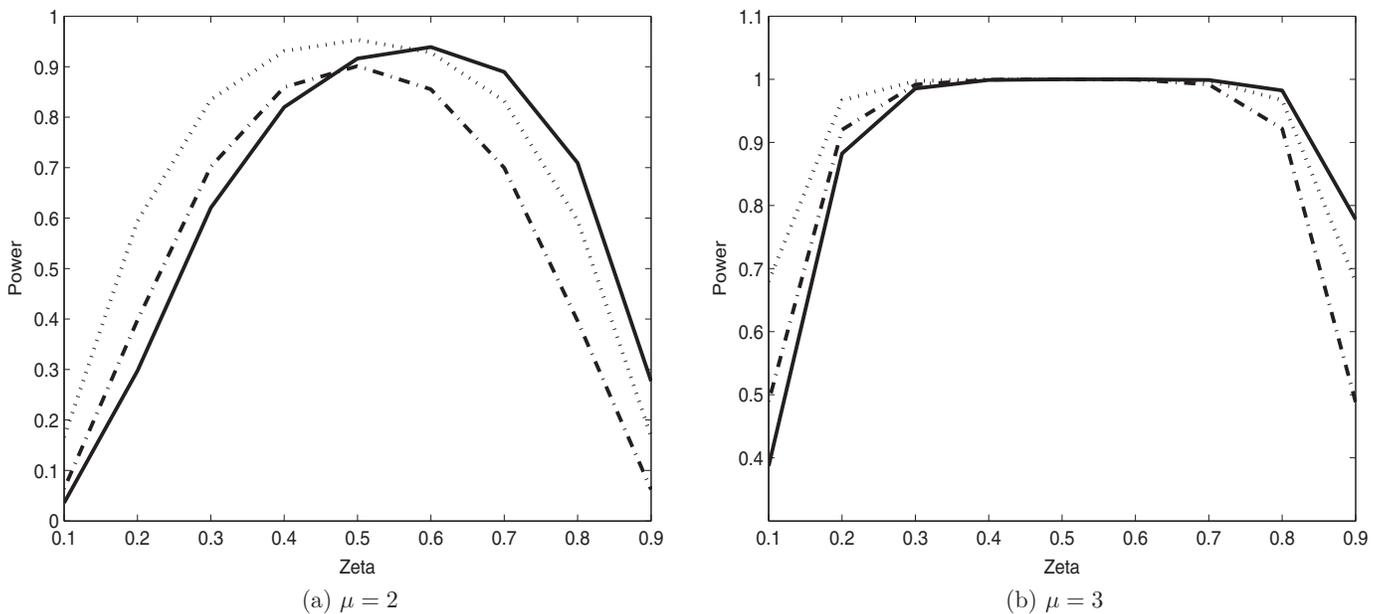
Figure 2. Power as a function of fraction ζ of sample size allocated to the primary study, for (a) $\mu_1 = \mu_2 = 2$, and (b) $\mu_1 = \mu_2 = 3$, for Procedure 4.1 with $w_1 = 1$ (solid), with $w_1 = 0.5$ (dotted), and for the BH procedure on the maximum of two studies p-values (dash-dotted) at level $q = 0.05$. The remaining parameters were $f_{00} = 0.9$, $f_{01} = 0.025$, $f_{10} = 0.025$, $f_{11} = 0.05$, sample size $N = 1000$, standard deviation $\sigma = 10$.
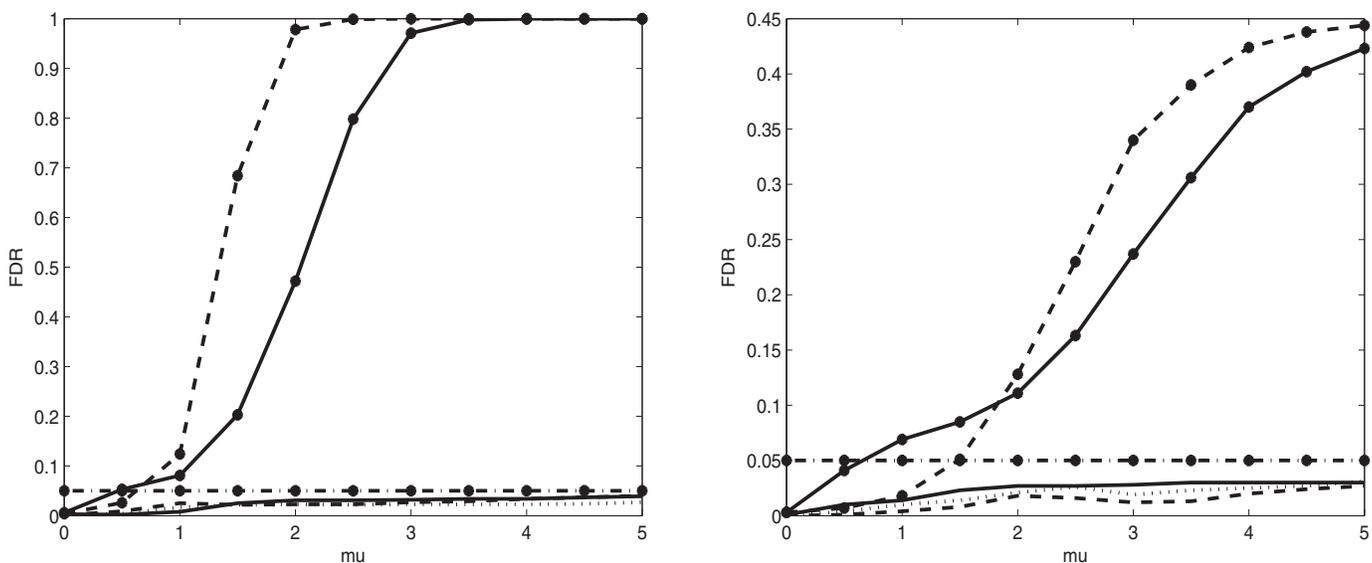
Figure 3. FDR versus $\mu = \mu_1 = \mu_2$ for $f_{01} = f_{10} = 0.5$ (left) and $f_{00} = 0.8$, $f_{01} = f_{10} = 0.1$ (right), for the following procedures at level $q = 0.05$: BH-1, BH-2 (solid with circles); BH-2, BH-1 (dashed with circles); Procedure 4.1 with $q_1 = 0.025$ and w_1 of 1 (solid), 0.5 (dotted), or 0 (dashed). The standard deviations were $\sigma_1 = 0.3$ and $\sigma_2 = 1$.

study p-values, where $k \in \{25, 30, \dots, 100\}$. The remaining parameters were: $f_{00} = 0.9$, $f_{01} = f_{10} = 0.025$, $f_{11} = 0.05$, $\sigma_1 = 0.5$, $\sigma_2 = 1$, $\mu_2 = 3$. For different values of μ_1 the optimal k is different, and using the BH procedure for selection is optimal for the entire range of μ_1.

7. DISCUSSION

In many research areas first a primary study is analyzed, then a follow-up study is analyzed with the goal to corroborate the

findings, or at least a subset of the findings, of the primary study. We suggested novel testing procedures for corroborating the evidence from a primary study in a follow-up study. We demonstrated their usefulness on a GWAS application. In the setting where there is no division of roles to a primary and a follow-up study, the simulations suggested that our novel Procedure 4.1 with $w_1 = 0.5$ is more powerful than the BH procedure on maximum p-values.

We proved that Procedures 3.2 and 4.1 control the FDR when the p-values are independent within each study and the selection
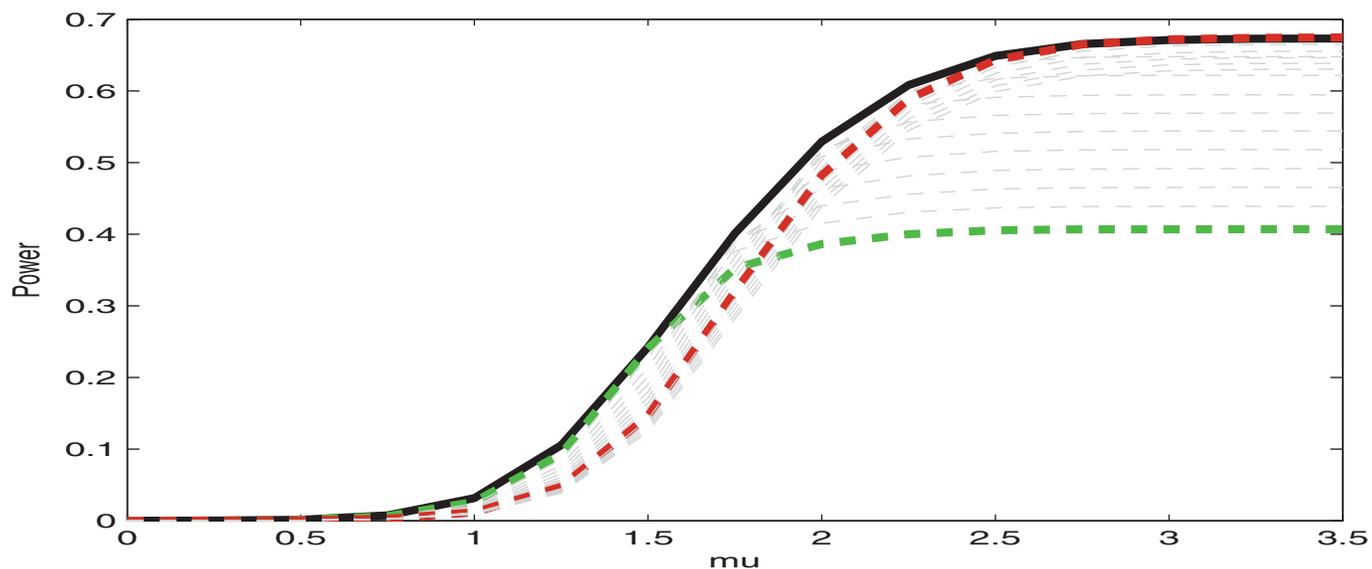
Figure 4. Power as a function of μ_1 for Procedure 4.1 with parameters $w_1 = 0.5$, $q_1 = 0.025$, $q = 0.05$ for the following selection rules: BH at level 0.0125 (solid black curve); selection of the hypotheses with k smallest primary study p-values, where $k = 25$ (dashed green curve), $k = 75$ (dashed red curve), $k \in \{30, 35, \dots, 100\}$ (dashed gray curves). The remaining parameters were: $f_{00} = 0.9$, $f_{01} = f_{10} = 0.025$, $f_{11} = 0.05$, $\sigma_1 = 0.5$, $\sigma_2 = 1$, $\mu_2 = 3$. The online version of this figure is in color.

rule is stable. However, the assumption of independence may not be realistic. Extensive simulations demonstrated that the BH procedure controls the FDR for many types of dependence encountered in practice (Yekutieli 2008). We conjecture that this robustness property carries over to Procedures 3.2 and 4.1, since Procedure 3.2 can be viewed as two-dimensional variant of the BH procedure. For simulated GWAS examples, the average FDP was below the nominal FDR level, suggesting that the procedures are indeed valid for the type of dependency that occurs in GWAS. More conservative variants of Procedure 3.2 were given in Theorem 3.3 and in Section 3 of the supplementary material, which guarantee that the FDR is controlled for arbitrary dependence among the primary study p-values, and dependence of type PRDS or arbitrary dependence among the follow-up study p-values. We demonstrated the usefulness of the variants suggested in Theorem 3.3 in Example 2 of Section 5. Out of the 36 replicability discoveries with Procedure 3.2, 23 discoveries passed the more stringent requirement that came with the added guarantee that the FDR is controlled for arbitrary dependence among the 635,547 p-values in the primary study.

Replicability analysis, as suggested in this article, requires that the investigators make several key design choices in addition to the error level q: the selection rule, q_1, and w_1 if two studies are available without division into primary and follow-up. The power of the procedure for replicability analysis varies with these choices. From our investigations, it appears reasonable in Procedure 3.2 to select hypotheses by BH at level q_1, and to set $w_1 = 0.5$ in Procedure 4.1 if the p-value distributions for false null hypotheses may be assumed to be similar in both studies. We gave some guidelines for choosing q_1 in specific settings, and more general guidelines are a topic for future research.

In replicability analysis, the primary study guides the design of the follow-up study by supplying the subset of hypotheses to be followed-up. Since the primary study also yields information on effect sizes, if it is assumed that the effect sizes are the same across studies, then this information may be used to determine the sample size needed to obtain good power in the follow-up study. However, this assumption may be unrealistic in applications such as GWAS, where the LD pattern varies across populations.

Finally, we saw that although Procedure 4.1 with parameters (w_1, q_1, q) is far less conservative than the BH procedure at level q on maximum p-values, it is still conservative. We proved that Procedure 4.1 with less conservative parameters $q_1' > q_1$ and $q' > q$ still controls the FDR at level q on the family of no replicability null hypotheses, if $|I_{00}|$ and $|I_{01}|$ were known. In future research we will consider estimates of these unknown parameters.

APPENDIX A: PROOF OF THEOREM 3.2

Let $q_2 = q - q_1$, and for each $j \in \{1, \ldots, m\}$, let $P_1^{(j)}$ and $P_2^{(j)}$ denote the vectors $P_1 = (P_{11}, \ldots, P_{1m})$ and $P_2 = (P_{21}, \ldots, P_{2m})$ with, respectively, P_{1j} and P_{2j} excluded. For $j \in \{1, \ldots, m\}$ arbitrary fixed, let $\mathcal{R}_1^{(j)}(P_1^{(j)}) \subseteq \{1, \ldots, j-1, j+1, \ldots, m\}$ be the subset of indices selected along with index j. Note that since the selection rule is stable, this subset is well defined. For any $j \in \{1, \ldots, m\}$ and given $P_1^{(j)}$, for

$i \in 1, \ldots, j-1, j+1, \ldots, m$ we define

$$T_i = \begin{cases} \max\left(\dfrac{m P_{1i}}{q_1}, \dfrac{(|\mathcal{R}_1^{(j)}(P_1^{(j)})| + 1)P_{2i}}{q_2} \right) & \text{if } i \in \mathcal{R}_1^{(j)}(P_1^{(j)}), \\ \infty & \text{otherwise.} \end{cases}$$

Let $T_{(1)} \leq \cdots \leq T_{(m-1)}$ be the sorted T-values, and $T_{(0)} = 0$. For $r = 1, \ldots, m$, we define $C_r^{(j)}$ as the event in which if $H_{\text{NR}, j}$ is rejected then, r hypotheses are rejected including $H_{\text{NR}, j}$. For Procedure 3.2, this amounts to:

$$C_r^{(j)} = \{(P_1^{(j)}, P_2^{(j)}) : T_{(r-1)} \leq r, T_{(r)} > r + 1, T_{(r+1)} \\ > r + 2, \ldots, T_{(m-1)} > m\}.$$

Note that given P_1, for $r > |\mathcal{R}_1|$, $C_r^{(j)} = \emptyset$, since exactly $|\mathcal{R}_1| - 1$ T_i's are finite.

Obviously, $C_r^{(j)}$ and $C_{r'}^{(j)}$ are disjoint events for any $r \neq r'$, and $\cup_{r=1}^m C_r^{(j)}$ is the entire space of $(P_1^{(j)}, P_2^{(j)})$. Let $I_0 = I_{01} \cup I_{00}$, R_j be the indicator of whether $H_{\text{NR}, j}$ was rejected for $j = 1, \ldots, m$, and $R = \sum_{j=1}^m R_j$. The FDR for the family of no replicability null hypotheses is

$$\text{FDR} = E\left(\frac{\sum_{j \in I_0} R_j}{\max(R, 1)} \right) + E\left(\frac{\sum_{j \in I_{10}} R_j}{\max(R, 1)} \right). \quad \text{(A.1)}$$

First, we find an upper bound for the first term of the sum in (A.1).

$$E\left(\frac{\sum_{j \in I_0} R_j}{\max(R, 1)} \right)$$

$$= \sum_{j \in I_0} \sum_{r=1}^m \frac{1}{r} \text{Pr}\left(j \in \mathcal{R}_1, P_{1j} \leq \frac{rq_1}{m}, P_{2j} \leq \frac{r(q - q_1)}{|\mathcal{R}_1|}, C_r^{(j)} \right)$$

$$\leq \sum_{j \in I_0} \sum_{r=1}^m \frac{1}{r} \text{Pr}\left(P_{1j} \leq \frac{rq_1}{m}, C_r^{(j)} \right)$$

$$= \sum_{j \in I_0} \sum_{r=1}^m \frac{1}{r} \text{Pr}\left(P_{1j} \leq \frac{rq_1}{m} \right) \text{Pr}\left(C_r^{(j)} \right) \quad \text{(A.2)}$$

$$\leq \frac{q_1}{m} \sum_{j \in I_0} \sum_{r=1}^m \text{Pr}\left(C_r^{(j)} \right) = \frac{|I_0|}{m} q_1. \quad \text{(A.3)}$$

The equality in (A.2) follows from the independence of the p-values. The inequality in (A.3) follows from the fact that for each $j \in I_0$, $\text{Pr}(P_{1j} \leq x) \leq x$ for all $x \in [0, 1]$. Finally, the equality in (A.3) follows from the fact that $\cup_{r=1}^m C_r^{(j)}$ is the entire sample space of $(P_1^{(j)}, P_2^{(j)})$, represented as a union of disjoint events.

Next, we find an upper bound for the second term of the sum in (A.1). Let $\mathcal{R}_1(p_1)$ be the set of selected indices using $P_1 = p_1$. Then $E(\sum_{j \in I_{10}} R_j / \max(R, 1) \mid P_1 = p_1)$ equals to

$$\sum_{j \in I_{10} \cap \mathcal{R}_1(p_1)} \sum_{r=1}^{|\mathcal{R}_1(p_1)|} \frac{1}{r} \mathbf{I}\left[p_{1j} \leq \frac{rq_1}{m} \right]$$

$$\times \text{Pr}\left(P_{2j} \leq \frac{rq_2}{|\mathcal{R}_1(p_1)|}, C_r^{(j)} \mid P_1 = p_1 \right)$$

$$\leq \sum_{j \in I_{10} \cap \mathcal{R}_1(p_1)} \sum_{r=1}^{|\mathcal{R}_1(p_1)|} \frac{1}{r} \text{Pr}\left(P_{2j} \leq \frac{rq_2}{|\mathcal{R}_1(p_1)|}, C_r^{(j)} \mid P_1 = p_1 \right) \quad \text{(A.4)}$$

$$= \sum_{j \in I_{10} \cap \mathcal{R}_1(p_1)} \sum_{r=1}^{|\mathcal{R}_1(p_1)|} \frac{1}{r} \text{Pr}\left(P_{2j} \leq \frac{rq_2}{|\mathcal{R}_1(p_1)|} \mid P_1 = p_1 \right)$$

$$\times \text{Pr}\left(C_r^{(j)} \mid P_1 = p_1 \right) \quad \text{(A.5)}$$

$$\leq \frac{q_2}{|\mathcal{R}_1(p_1)|} \sum_{j \in I_{10} \cap \mathcal{R}_1(p_1)} \sum_{r=1}^{|\mathcal{R}_1(p_1)|} \text{Pr}\left(C_r^{(j)} \mid P_1 = p_1 \right) \quad \text{(A.6)}$$

$$= \frac{q_2}{|\mathcal{R}_1(p_1)|} |I_{10} \cap \mathcal{R}_1(p_1)|. \quad \text{(A.7)}$$

The equality in (A.5) follows from the fact that P_{2j}, $P_2^{(j)}$, P_1 are independent, since then $C_r^{(j)}$ and the event $\{P_{2j} \le rq_2/|\mathcal{R}_1(p_1)|\}$ are conditionally independent. The inequality in (A.6) follows from the independence of the p-values across the studies and the fact that for each $j \in I_{10}$, $\Pr(P_{2j} \le x) \le x$ for all $x \in [0, 1]$. The equality in (A.7) follows from the fact that $\cup_{r=1}^{|\mathcal{R}_1(p_1)|} C_r^{(j)}$ is a union of disjoint events, and $\Pr(\cup_{r=1}^{|\mathcal{R}_1(p_1)|} C_r^{(j)} \mid P_1 = p_1) = 1$.

It follows from (A.7) that $E(\sum_{j \in I_{10}} R_j / \max(R, 1)) \le q_2$. Using this fact and the bound (A.3) for the first term of (A.1), we obtain

$$\text{FDR} \le \frac{|I_0|}{m} q_1 + (q - q_1) \le q_1 + (q - q_1) = q.$$

APPENDIX B: PROOF FOR FDR CONTROL OF THE ORACLE PROCEDURE 3.2

Let us now prove that under the assumption that the p-values are independent, Procedure 3.2 at levels $(q', 2q')$ controls the FDR at level $|I_{00}|(q')^2/m + (|I_{01}|/m + 1)q'$. Returning to the proof of Theorem 3.2, note that (A.1) can be rewritten as follows:

$$\text{FDR} = E\left(\frac{\sum_{j \in I_{00}} R_j}{\max(R, 1)}\right) + E\left(\frac{\sum_{j \in I_{01}} R_j}{\max(R, 1)}\right) + E\left(\frac{\sum_{j \in I_{10}} R_j}{\max(R, 1)}\right). \tag{B.1}$$

We will now give an upper bound for each term of the sum in (B.1). First,

$$E\left(\frac{\sum_{j \in I_{00}} R_j}{\max(R, 1)}\right)$$

$$= \sum_{j \in I_{00}} \sum_{r=1}^{m} \frac{1}{r} \Pr\left(j \in \mathcal{R}_1, P_{1j} \le \frac{rq'}{m}, P_{2j} \le \frac{rq'}{|\mathcal{R}_1|}, C_r^{(j)}\right)$$

$$\le \sum_{j \in I_{00}} \sum_{r=1}^{m} \frac{1}{r} \Pr\left(P_{1j} \le \frac{rq'}{m}, P_{2j} \le q', C_r^{(j)}\right)$$

$$\le \frac{(q')^2}{m} \sum_{j \in I_{00}} \sum_{r=1}^{m} \Pr\left(C_r^{(j)}\right) = \frac{|I_{00}|}{m}(q')^2. \tag{B.2}$$

The inequality in (B.2) follows from the facts that for each $j \in I_{00}$, P_{1j} and P_{2j} are independent, and $\Pr(P_{ij} \le x) \le x$ for all $x \in [0, 1]$ and $i = 1, 2$. The equality in (B.2) follows from the explanation of the equality in (A.3).

Second, replacing I_0 by I_{01} and $|I_0|$ by $|I_{01}|$ in the arguments that led to (A.3), we obtain

$$E\left(\frac{\sum_{j \in I_{01}} R_j}{\max(R, 1)}\right) \le \frac{|I_{01}|}{m} q'. \tag{B.3}$$

Finally, using (A.7) in the proof of Theorem 3.2 we obtain that the third term of the sum in (B.1) is bounded by $q_2 = 2q' - q' = q'$. Using this upper bound, together with the bounds for the first two terms derived in (B.2) and (B.3), we obtain

$$\text{FDR} \le \frac{|I_{00}|}{m}(q')^2 + \frac{|I_{01}|}{m}q' + q' = \frac{|I_{00}|}{m}(q')^2 + \left(\frac{|I_{01}|}{m} + 1\right)q'.$$

It follows that if $|I_{00}|$ and $|I_{01}|$ were known, one could guarantee FDR control at level q on the family of no replicability null hypotheses by applying Procedure 3.2 at levels $(q', 2q')$, where q' is the solution to $|I_{00}|(q')^2/m + (|I_{01}|/m + 1)q' = q$.

APPENDIX C: PROOF OF THEOREM 4.1

Let $V_{12} = \sum_{j \in I_{00} \cup I_{01} \cup I_{10}} \mathbf{I}[j \in \mathcal{R}_{12,w_1q}]$ and $R_{12} = |\mathcal{R}_{12,w_1q}|$ denote the number of erroneously rejected and the total number of rejected no replicability null hypotheses by Procedure 3.2 at level w_1q with study one as the primary study and study two as the follow-up study. Similarly, let $V_{21} = \sum_{j \in I_{00} \cup I_{01} \cup I_{10}} \mathbf{I}[j \in \mathcal{R}_{21,(1-w_1)q}]$ and $R_{21} = |\mathcal{R}_{21,(1-w_1)q}|$ denote the number of erroneously rejected and the total number of rejected no replicability null hypotheses by Procedure 3.2 at level $(1 - w_1)q$ with study two as the primary study and study one as the follow-up study. Define $\mathcal{R}_s = \mathcal{R}_{12,w_1q} \cup \mathcal{R}_{21,(1-w_1)q}$, the indices of the no replicability null hypotheses rejected by Procedure 4.1. Let $V_s = \sum_{j \in I_{00} \cup I_{01} \cup I_{10}} \mathbf{I}[j \in \mathcal{R}_s]$ and $R_s = |\mathcal{R}_s|$, the number of erroneously rejected and the total number of rejected no replicability null hypotheses by Procedure 4.1.

Note that $V_s \le V_{12} + V_{21}$. Therefore,

$$\text{FDR} = E\left(\frac{V_s}{\max(R_s, 1)}\right) \le E\left(\frac{V_{12}}{\max(R_s, 1)}\right) + E\left(\frac{V_{21}}{\max(R_s, 1)}\right). \tag{C.1}$$

In addition, note that $\max(R_s, 1) \ge \max(R_{12}, 1)$ and $\max(R_s, 1) \ge \max(R_{21}, 1)$. Using these facts and (C.1), we obtain

$$\text{FDR} = E\left(\frac{V_s}{\max(R_s, 1)}\right) \le E\left(\frac{V_{12}}{\max(R_{12}, 1)}\right) + E\left(\frac{V_{21}}{\max(R_{21}, 1)}\right)$$
$$\le w_1q + (1 - w_1)q = q,$$

where the last inequality follows from Theorem 3.2.

APPENDIX D: TABLE OF RESULTS FOR GWAS OF CROHN'S DISEASE

Table D.1. Replicability analysis for Example 2 in Section 5: GWAS of Crohn's disease. The number of SNPs in the primary study was 635,547, and 126 SNPs were followed-up. The 36 discoveries by Procedure 3.2 with parameters $(q_1, q) = (0.04, 0.05)$ are listed according to the adjusted p-values. The primary and follow-up studies p-values are given in columns 4 and 5; the adjusted p-values for $c = 0.8$ are given in column 6 for Procedure 3.2, and in column 7 for the modification of item 1 in Theorem 3.3

Index	Chromosome	Position	p_1	p_2	$p_{\text{FDR}}^{\text{REPadj}}$	$\tilde{p}_{\text{FDR}}^{\text{REPadj}}$
1	1	67417979	3.19e-34	1.5e-36	2.53e-28	3.53e-27
2	1	67414547	5.05e-36	3.1e-29	9.69e-27	9.69e-27
3	1	67387537	1.35e-24	5.62e-17	1.17e-14	1.17e-14
4	2	233962410	5.66e-21	7.67e-14	1.2e-11	1.2e-11
5	10	64108492	9.51e-12	1.61e-10	1.51e-06	1.5e-05
6	5	40428485	2.51e-22	2.79e-08	2.84e-06	3.31e-06
7	5	40437266	2.26e-22	3.18e-08	2.84e-06	3.31e-06
8	10	101281583	8.53e-11	1.69e-07	1.32e-05	7.74e-05
9	18	12769947	5.95e-12	2.41e-07	1.61e-05	1.88e-05
10	5	150239060	3.18e-11	2.57e-07	1.61e-05	3.91e-05
11	10	101282445	9.09e-11	3.1e-07	1.76e-05	7.74e-05
12	5	150203580	4.09e-11	7.47e-07	3.89e-05	4.67e-05
13	18	12799340	3.27e-11	1.23e-06	5.91e-05	6.99e-05
14	5	131798704	2.29e-09	3.52e-11	0.00013	0.00169
15	5	158747111	4.4e-09	3.66e-06	0.000233	0.00305
16	2	233965368	1.28e-21	3.66e-05	0.00143	0.00163
17	13	43355925	8.04e-08	1.33e-07	0.00376	0.0469
18	12	39104262	8.95e-08	6.55e-05	0.00395	0.0496
19	3	49676987	9.47e-08	2.24e-06	0.00396	0.0499
20	3	49696536	1.08e-07	5.64e-07	0.00429	0.0544
21	12	38888207	6.64e-08	0.000165	0.00491	0.0433
22	6	167408399	1.65e-07	3.26e-07	0.00596	0.0731
23	9	114645994	1.96e-07	6.58e-05	0.00677	0.0768
24	6	20836710	1.26e-07	0.000278	0.00724	0.0607
25	1	169593891	2.01e-07	0.000321	0.00802	0.0768
26	1	197667523	3.41e-07	2.34e-06	0.01	0.111
27	9	4971602	3.4e-07	0.00043	0.01	0.111
28	1	157665119	1.75e-07	0.000481	0.0107	0.0745
29	11	75978964	7.16e-08	0.000732	0.0158	0.044
30	20	61798026	7.6e-07	0.000138	0.0201	0.234
31	6	167405736	1.65e-07	0.00121	0.0241	0.0731
32	1	197691964	9.69e-07	1e-04	0.0241	0.29
33	17	35294289	1.06e-06	0.000292	0.0255	0.308
34	8	126603853	1.9e-06	0.000182	0.0431	0.457
35	6	106541962	1.85e-06	7.7e-06	0.0431	0.457
36	9	4978761	1.96e-06	0.00162	0.0433	0.462

SUPPLEMENTARY MATERIALS

Proof of Theorem 3.3, as well as additional theoretical results under dependency. An example that shows the advantage of Procedure 3.1 over an FWER controlling procedure on the maximum p-values is also presented.

[Received July 2012. Revised May 2013.]

REFERENCES

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J., NIDDK IBD Genetics Consortium, Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J. P., de Vos, M., Vermeire, S., Louis, E., Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium , Cardon, L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N. J., Onnie, C. M., Fisher, S. A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M., and Daly, M. J. (2008), "Genome-Wide Association Defines More Than 30 Distinct Susceptibility Loci for Crohn's Disease," *Nature Genetics*, 40, 955–962. [1484]

Benjamini, Y., and Heller, R. (2008), "Screening for Partial Conjunction Hypotheses," *Biometrics*, 64, 1215–1222. [1481]

Benjamini, Y., Heller, R., and Yekutieli, D. (2009), "Selective Inference in Complex Research," *Philosophical Transactions of the Royal Society A*, 267, 1–17. [1481]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate—A Practical and Powerful Approach to Multiple Testing," *Journal of The Royal Statistical Society,* Series B, 57, 289–300. [1481]

——— (2000), "On the Adaptive Control of the False Discovery Fate in Multiple Testing With Independent Statistics," *Journal of Educational and Behavioral Statistics*, 25, 60–83. [1482]

Benjamini, Y., Krieger, M., and Yekutieli, D. (2006), "Adaptive Linear Step-Up False Discovery Rate Controlling Procedures," *Biometrika*, 93, 491–507. [1482]

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188. [1483]

———— (2005), "Quantitative Trait Loci Analysis Using the False Discovery Rate," *Genetics*, 171, 783–790. [1480,1483]

Bis, J. C., DeCarli, C., Smith, A. V., van der Lijn, F., Crivello, F., Fornage, M., Debette, S., Shulman, J. M., Schmidt, H., Srikanth, V., Schuur, M., Yu, L., Choi, S. H., Sigurdsson, S., Verhaaren, B. F., DeStefano, A. L., Lambert, J. C., Jack, C. R., Jr., Struchalin, M., Stankovich, J., Ibrahim-Verbaas, C. A., Fleischman, D., Zijdenbos, A., den Heijer, T., Mazoyer, B., Coker, L. H., Enzinger, C., Danoy, P., Amin, N., Arfanakis, K., van Buchem, M. A., de Bruijn, R. F., Beiser, A., Dufouil, C., Huang, J., Cavalieri, M., Thomson, R., Niessen, W. J., Chibnik, L. B., Gislason, G. K., Hofman, A., Pikula, A., Amouyel, P., Freeman, K. B., Phan, T. G., Oostra, B. A., Stein, J. L., Medland, S. E., Vasquez, A. A., Hibar, D. P., Wright, M. J., Franke, B., Martin, N. G., Thompson, P. M.; Enhancing Neuro Imaging Genetics through Meta-Analysis Consortium, Nalls, M. A., Uitterlinden, A. G., Au, R., Elbaz, A., Beare, R. J., van Swieten, J. C., Lopez, O. L., Harris, T. B., Chouraki, V., Breteler, M. M., De Jager, P. L., Becker, J. T., Vernooij, M. W., Knopman, D., Fazekas, F., Wolf, P. A., van der Lugt, A., Gudnason, V., Longstreth, W. T., Jr, Brown, M. A., Bennett, D. A., van Duijn, C. M., Mosley, T. H., Schmidt, R., Tzourio, C., Launer, L. J., Ikram, M. A., Seshadri, S.; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. (2012), "Common Variants at 12q14 and 12q24 are Associated With Hippocampal Volume," *Nature Genetics*, 44, 545–551. [1480,1481,1484]

Blanchard, G., and Roquain, E. (2009), "Adaptive False Discovery Rate Control Under Independence and Dependence," *Journal of Machine Learning Research*, 10, 2837–2871. [1482]

Hedges, L., and Olkin, I. (1985), *Statistical Methods for Meta-Analysis*, London: Academic Press. [1480,1481]

Kraft, P., Zeggini, E., and Ioannidis, J. (2009), "Replication in Genome-Wide Association Studies," *Statistical Science*, 24, 561–573. [1480]

Lander, E., and Kruglyak, L. (1995), "Genetic Dissection of Complex Traits: Guidelines for Interpreting and Reporting Linkage Results," *Nature Genetics*, 11, 241–247. [1480,1483]

Loughin, T. (2004), "A Systematic Comparison of Methods for Combining p-Values From Independent Tests," *Computational Statistics and Data Analysis*, 47, 467–485. [1485]

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003), "Identifying Differentially Expressed Genes Using False Discovery Rate Controlling Procedures," *Bioinformatics*, 19, 368–375. [1482]

Rosenbaum, P. (2001), "Replicating Effects and Biases," *The American Statistician*, 55, 223–227. [1480]

Simes, R. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 73, 751–754. [1483]

Skol, A., Scott, L., Abecasis, G., and Boehnke, M. (2006), "Joint Analysis is More Efficient Than Replication-Based Analysis for Two-Stage Genome-Wide Association Studies," *Nature Genetics*, 38, 209–213. [1480]

Storey, J., Taylor, J., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society,* Series B, 66, 187–205. [1482]

Storey, J., and Tibshirani, R. (2003), "Statistical Significance for Genomewide Studies," *Proceedings of the National Academy of Sciences*, 100, 9440–9445. [1482]

Su, Z., Marchini, J., and Donnelly, P. (2011), "Hapgen2: Simulation of Multiple Disease Snps," *Bioinformatics*, 27, 2304–2305. [1481,1485]

The International HapMap Consortium (2003), "The International Hapmap Project," *Nature*, 426, 789–796. [1485]

Yekutieli, D. (2008), Comments on "Control of the False Discovery Rate Under Dependence Using the Bootstrap and Subsampling," *Test*, 17, 458–460. [1489]

Zeggini, E., Weedon, M., Lindgren, C., Frayling, T., Elliott, K., Lango, H., Timpson, N., Perry, J., and Rayner, N. (2007), "Replication of Genome-Wide Association Signals in Uk Samples Reveals Risk Loci for Type 2 Diabetes," *Science*, 316, 1336–1341. [1480]