Estimating the evidence of replicability in "omics" research from follow-up studies

October 3, 2013

Ruth Heller
Department of Statistics and Operations Research, Tel-Aviv university, Tel-Aviv, Israel. E-mail: ruheller@post.tau.ac.il
Marina Bogomolov
Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Haifa, Israel. E-mail: marinabo@tx.technion.ac.il
Yoav Benjamini
Department of Statistics and Operations Research, Tel-Aviv university, Tel-Aviv, Israel. E-mail: ybenja@post.tau.ac.il

Abstract

In "omics" research primary studies are often followed by follow-up studies on promising findings. Informally, reporting the p-values of promising findings in the primary and the follow-up studies gives a sense of the replicability of the findings. We offer a formal statistical approach to measure the evidence of replicability of findings from the primary study to the follow-up study, that exploits the fact that most hypotheses examined in the primary study are null. For each of the findings in the follow-up study we compute the lowest false discovery rate at which the finding is called replicated, which we term the r-value.

We are concerned with situations in which many features are scanned for their statistical significance in a primary study. These features can be single nucleotide polymorphisms (SNPs) associations with disease, gene expressions differences, pathways enrichments, protein-protein interactions, etc. Interesting features are selected for follow-up, and only the selected ones are tested in a follow-up study.

This approach addresses two goals. The first goal is to increase the number of cases in order to increase the power to detect a feature, at a lower cost. The second goal is to address the basic dogma of science that a finding is more convincingly a true

finding if it is replicated in at least one more study, and the requirement becomes essential with the use of high throughput methods. The paramount importance of having replicated findings is well recognized in genomic research, see [1]. In particular, this is so in genome-wide association studies (GWAS), see [2] and [3]. As noted in [4], the anticipated effects for common variants in GWAS 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 argue that it is important to see the association in other studies using similar, but not identical, study base and methods.

A typical table of results in GWAS reports the p-values in the primary and follow-up study, side by side, as well as the meta-analysis p-values, for the SNPs with the smallest meta analysis p-values. Table 2 columns 1-6 is an example of such a table of results [5]. However, a small meta-analysis p-value addresses only the first goal of increased power, and is not evidence towards replicability, since it tests only the null hypothesis of no signal in both studies. As a result, a strong signal in one of the studies (with p-value close to zero) is enough to declare the meta-analysis finding as highly significant as well. Here we address the testing of no signal in one or more studies, which yields the statistical significance of the replicability claim. (Replicability is sometimes referred to as reproducibility, but see [6].)

The r-value for replicability. Even for a single study raw p-values are not useful enough, because few features are being selected from the many scanned. Multiple testing methods are widely employed to adjust for the effect of selection, either by controlling the probability of erroneously selecting even a single feature (FWER), or by controlling the false discovery rate (FDR). The concern regarding the selected claims of replicability is even greater, because the selection takes place both after the primary study and after the follow-up study. We therefore propose to report the smallest FDR level for which the feature has been significantly replicated from the primary study to the follow-up study. The smaller this value, the stronger the evidence in favor of the replicability of the finding. In accordance with the q-value [7], that is similar to the FDR adjusted p-value [8], we coin the term r-value for the FDR-adjusted evidence towards replicability of the finding. It can then be compared to any desired level of FDR in the same way that a p-value, or a q-value, is compared to the desired level α. We thus suggest to complement the table of results on the SNPs followed-up in this typical GWAS example with the r-values in column 7. They are all below 0.05, concurring with the main replicability findings of [5]. The ranking of r-values is different than the ranking of the meta-analysis p-values, indicating the novelty of the added information. The result of the somewhat more complicated example to be discussed shortly are given in Table 3, where the difference between the meta-analysis conclusions and the replicability conclusions are more dramatic.

Previous works that discuss replicability of results in multiple studies with no division into primary and follow-up studies include [9], that suggest a method based on relative ranking of the p-values to control their "irreproducible discovery rate";

[10], that suggest a list-intersection test to compare top ranked gene lists from multiple studies in order to discover the common significant set of genes; and [11], that suggest an empirical Bayes approach for discovering whether results have been replicated across studies. In this work, we address the important setting where there is a clear division between the primary and the follow-up study, and the features selected for follow-up are based on the data from the primary study. A formal procedure suggested in [12] was to use the maximum of the two studies p-values for each feature as a way to test its replicability of significance. [13] suggested a more powerful way to discover the replicable findings when the selected features for follow-up are a small fraction of those tested in the primary study. Here we suggest a generalization of the method of [13], which offers further power gain in the typical situation in "omics" research where most of the hypotheses examined in the primary study are true null hypotheses.

1 Proposed Method and Results

The family of m hypotheses examined in the primary study, indexed by $I = \{1, \ldots, m\}$, may be divided into four sub-families with the following indices: I_{00}, for the hypotheses that are null in both studies; I_{01}, for the hypotheses that are null in the primary study only; I_{10}, for the hypotheses that are null in the follow-up study only; and I_{11}, for the hypotheses that are non-null in both studies. 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, and $R_1 = |\mathcal{R}_1|$ be their number. Suppose R replicability claims are made by an analysis. Denoting by R_{ij} the number of replicability claims from sub-family I_{ij}, R_{11} is the number of true replicability claims, and $R - R_{11} = R_{00} + R_{01} + R_{10}$ is the number of false replicability claims.

We are interested in controlling the FDR for replicability analysis,

$$FDR = E\left(\frac{R_{00} + R_{01} + R_{10}}{\max(R, 1)}\right),$$

which is the expected proportion of false replicability claims among all those called replicated.

Let $f_{0j} = \frac{|I_{0j}|}{m}, j \in \{0, 1\}$, and $f_{10} = \frac{|I_{10}|}{m}$. We suggest a procedure for replicability analysis, that can exploit the fact that f_{00} is typically closer to one than to zero. Although f_{00} is not estimable from the data, since we only observe data for m SNPs in the primary study, we can give a conservative guess for a lower bound on f_{00}, call it l_{00}. For example, for GWAS on the whole genome, $l_{00} = 0.8$ is conservative in typical applications.

Let p_{1j} be the primary study p-value for hypothesis j, $j \in \{1, \ldots, m\}$, and p_{2j} be the follow-up study p-value for hypothesis j, $j \in \mathcal{R}_1$. We suggest the following procedure for replicability analysis.

Procedure 1.1 *Replicability analysis with parameters* (l_{00}, c_2, q)*, where* $0 \leq l_{00} < 1$*,* $0 < c_2 < 1$*, and* $0 < q < 1$*:*

1. *Let* $c_1 = \frac{1-c_2}{1-l_{00}(1-c_2 q)}$.
2. *For every feature* $j \in \mathcal{R}_1$ *compute the following E-values*

$$E_j = \max\left(\frac{mp_{1j}}{c_1}, \frac{R_1 p_{2j}}{c_2}\right), j \in \mathcal{R}_1,$$

3. *The r-value for feature* $i \in \mathcal{R}_1$ *is*

$$r_i = \min_{E_j \geq E_i, j \in \mathcal{R}_1} \frac{E_j}{rank(E_j)},$$

   where $rank(E_j)$ *is the rank of* E_j *among the* R_1 *E-values, with maximum value for ties.*

4. *Declare the set of findings of features with* $r_i \leq q$ *as replicated.*

The FDR level of Procedure 1.1 is at most

$$FDR \leq f_{00} c_1 c_2 q^2 + f_{01} c_1 q + E\left(\frac{|I_{10} \cap \mathcal{R}_1|}{|\mathcal{R}_1|}\right) c_2 q, \tag{1}$$

if the p-values in the follow-up study are independent or have positive dependency of type PRDS, and either one of the following (1) the p-values in the primary study are independent, or (2) the p-values in the primary study have arbitrary dependence, but only hypotheses with primary study p-values below $c_1 q/m$ are considered for follow-up, see Supporting Information (SI) for a proof. The two last terms of the upper bound in equation [1] are reached when $p_{2j} \approx 0$ for $j \in I_{01}$ and $p_{1j} \approx 0$ for $j \in I_{10}$.

Since the upper bound [1] is at most

$$f_{00} c_1 c_2 q^2 + f_{01} c_1 q + c_2 q,$$

if the constants (l_{00}, c_2) satisfy the inequality

$$f_{00} c_1 c_2 q + f_{01} c_1 + c_2 \leq 1,$$

then Procedure 1.1 controls the FDR for replicability analysis at level q. This inequality holds for any choice of (l_{00}, c_2) in Procedure 1.1 that satisfies the relationship

$$l_{00} \leq \frac{1 - f_{01} - f_{00} c_2 q}{1 - c_2 q}.$$

Unfortunately, f_{00} and f_{01} are not known. If the guess for l_{00} is indeed conservative, i.e. $l_{00} \leq f_{00}$, then the above inequality holds since $f_{00} \leq 1 - f_{01}$. Thus, for any value

$l_{00} \leq f_{00}$ and $c_2 \in (0,1)$, Procedure 1.1 controls the FDR for replicability analysis at level q.

For q small, step 1 of Procedure 1.1 leads to $c_1 \approx \frac{1-c_2}{1-l_{00}}$. Using this approximation, the r-values do not depend on q, and reporting the r-values is useful for quantifying the evidence of replicability as well as for selecting the replicability claims: the replicability claims at level q are all indices with r-values below q. For the choice $l_{00} = 0.8$, which is a conservative lower bound for typical GWAS on the whole genome, $c_1 \approx 5(1-c_2)$ for q small. This parameter choice results in r-values that are up to five times smaller than the r-values from Procedure 1.1 with $l_{00} = 0$.

For $l_{00} = 0$, Procedure 1.1 coincides with Procedure 3.2 in [13]. It is easy to see that Procedure 1.1 with parameters (l_{00}, c_2, q), where $l_{00} > 0$, will have at least as many replicability claims as Procedure 3.2 in [13]. We show in examples and simulations that the power of Procedure 1.1 increases with l_{00}, and can lead to many more discoveries than with Procedure 3.2 in [13], while maintaining FDR control.

1.1 Parameter choices

How should c_2 be chosen? In simulations, detailed in the SI, we observed that Procedure 1.1 with optimal c_2 for a given l_{00} that maximizes power, has only a small gain in power over the choice $c_2 = 0.5$. Table 1 summarizes the results for the following specific simulation setting. We considered $m = 1000$ SNPs, out of which $f_{00} = 0.9$ contained no signal, $f_{01} = 0.025$ contained signal only in the follow-up study, $f_{10} = 0.025$ contained signal only in the primary study, and $f_{11} = 0.05$ had signal in both studies. The power to detect the signal in the primary study was set to be $\pi_1 = 0.1$ for a threshold of $0.05/m$, and the power to detect the signal in the follow-up study was set to be $\pi_2 \in \{0.8, 0.5, 0.2\}$ for a threshold of $0.05/R_1$. The selection rule for follow-up was the Benjamini-Hochberg (BH) procedure [14] at level $c_1 q$ on the primary study p-values, with $q = 0.05$. Table 1 shows, as we expect, that the power increases with l_{00} as well as with π_2. Moreover, the difference in power between the best choice of c_2 (which is unknown in advance), and our suggestion $c_2 = 0.5$, is small. The gain in power of using Procedure 1.1 with $l_{00} > 0$ over Procedure 1.1 with $l_{00} = 0$ can be large, as can be seen by the comparison of columns 8-9 with columns 3, 5, and 7 in Table 1. In the SI, Figure S1 shows the average power and the power for at least one true replicability discovery as a function of c_2. In all simulations, in Procedure 1.1 step 1 we used the approximation $c_1 \approx \frac{1-c_2}{1-l_{00}}$, so there was no need to prefix q for the r-value computation.

Our simulations mimic the typical setting in GWAS on the whole genome, where SNPs that are associated with the phenotype have typically low power (0.1 in the above simulations) to pass the severe Bonferroni threshold of the large number of hypotheses examined in the primary study, yet the power to pass the far less severe Bonferroni threshold of the few dozen hypotheses examined in the follow-up study

Table 1: The estimated average power of Procedure 1.1 with parameters $(l_{00}, c_2, 0.05)$, where c_2 is the optimal choice among the values in $\{0.05, 0.1, \ldots, 0.95\}$ for $l_{00} = 0.5$ (column 2), $l_{00} = 0.8$ (column 4), $l_{00} = 0.9$ (column 6), the optimal value of c_2 is given in the row below; $c_2 = 0.5$, for $l_{00} \in \{0.5, 0.8, 0.9\}$ (columns 3, 5, 7); $l_{00} = 0$ and $c_2 \in \{0.5, 0.2\}$ (columns 8, 9) in a configuration $f_{00} = 0.9, f_{01} = f_{10} = 0.025, f_{11} = 0.05$. The number of hypotheses examined in the primary study is 1000. The signal to noise ratios for the primary study and the follow-up study, μ_1/σ_1 and μ_2/σ_2, respectively, are taken according to the requirement that the power of Bonferroni procedure at level 0.05 in the primary study is π_1, and in the follow-up study is π_2 (given in the first column). The standard errors were of the order of 10^{-3} for all the estimates.

(π_1, π_2)	Optimal for $l_{00} = 0.5$	$l_{00} = 0.5$ $c_2 = 0.5$	Optimal for $l_{00} = 0.8$	$l_{00} = 0.8$ $c_2 = 0.5$	Optimal for $l_{00} = 0.9$	$l_{00} = 0.9$ $c_2 = 0.5$	$l_{00} = 0$ $c_2 = 0.5$	$l_{00} = 0$ $c_2 = 0.2$
(0.1, 0.8)	0.2959 $c_2 = 0.2$	0.2541	0.4522 $c_2 = 0.2$	0.3990	0.5803 $c_2 = 0.15$	0.5271	0.1728	0.2018
(0.1, 0.5)	0.1776 $c_2 = 0.4$	0.1696	0.3031 $c_2 = 0.35$	0.2948	0.4172 $c_2 = 0.45$	0.4037	0.1050	0.1013
(0.1, 0.2)	0.0467 $c_2 = 0.55$	0.0450	0.0838 $c_2 = 0.55$	0.0831	0.1319 $c_2 = 0.5$	0.1319	0.0254	0.0154

is greater (0.2, 0.5 or 0.8 in the above simulations). Therefore, for GWAS on the whole genome, we recommend setting $c_2 = 0.5$. We suggest computing the r-values by Procedure 1.1 with parameters $(l_{00}, c_2, q) = (0.8, 0.5, q)$.

In the SI we further show realistic GWAS simulations that preserve the dependency across p-values in each study. The FDR is controlled below level 0.05 for Procedure 1.1 with parameters $(l_{00}, c_2, q) = (l_{00}, 0.5, 0.05)$, suggesting that this procedure is valid for the type of dependency that occurs in GWAS. Since Procedure 1.1 can be viewed as a two dimensional variant of the BH procedure, and the BH procedure is known to be robust to many types of dependencies, we conjecture that Procedure 1.1 with parameters $(l_{00}, c_2, q) = (l_{00}, 0.5, q)$, where $l_{00} \leq f_{00}$, controls the FDR at the nominal level q for most types of dependencies that occur in practice, even if hypotheses with primary study p-values above $c_1 q/m$ are followed-up.

1.2 Examples

We consider three recent articles reporting GWAS, where hundreds of thousands of SNPs are examined in the primary studies, and only a small fraction of these SNPs are examined in the follow-up studies. We quantify the evidence of replicability of associations for the SNPs followed-up in terms of the r-values computed with $c_2 = 0.5$ and three values of the lower bound: $l_{00} \in \{0.8, 0.5, 0\}$. In Procedure 1.1 step 1 we used the approximation $c_1 \approx \frac{1-c_2}{1-l_{00}}$, so there was no need to prefix q for the r-value computation. The examples differ in design, and in the selection rules for forwarding SNPs for follow-up. In the first example, the simplest design is considered where there is one primary study and one follow-up study, few dozen SNPs are followed up, and only a handful have r-values below 0.05. In the second example, the primary study

is a meta-analysis of three studies, more than a hundred hypotheses are followed-up, and few dozen SNPs have r-values below 0.05. In the third example, there were three stages: first a primary study, then a follow-up study, and then an additional follow-up study that was based on the first follow-up study.

Our first example is GWAS of IgA nephropathy in Han Chinese. To discover association between SNPs and IgA nephropathy, [5] measured 444882 SNPs in 1523 cases from southern China, and 4276 controls from southern China, from northern China, and from Singapore with the same ancestral origin. For follow-up, 61 SNPs were measured in two studies: 1402 cases and 1716 controls from northern China, and 1301 cases and 1748 controls from southern China. The 61 SNPs selected for follow-up had primary study p-values below 10^{-5}. Table 2 shows the seven SNPs with smallest meta-analysis p-values, out of the 61 SNPs followed up. All the seven findings discovered to be associated with the disease had r-values below 0.05 for $l_{00} = 0.8$. These results suggest that for these seven SNPs, the associations with the disease have been replicated. The seven SNPs clearly stand out from the remaining 54 SNPs followed-up, that have r-values of one, see Table S2 in the SI. If the researcher is willing to assume only a lower bound of 0.5 or of zero for f_{00}, then the r-values are larger than with $l_{00} = 0.8$. Table S2 in the SI shows that with $l_{00} = 0.5$, six SNPs had r-values below 0.05, and with $l_{00} = 0$, five SNPs had r-values below 0.05. The ranking of evidence towards replicability was different than the ranking of evidence towards association. The SNP in row number one has by far the smallest meta-analysis p-value, but in terms of evidence towards replicability it is less pronounced than the evidence of the SNP in the third row, since while both have very convincing evidence towards association in the follow-up study that examined simultaneously only 61 SNPs, the evidence of association in the primary study was more pronounced for the SNP in the third row than for the SNP in the first row.

Table 2: Replicability analysis for FDR control for the study of [5]: GWAS of IgA nephropathy in Han Chinese. The number of SNPs in the primary study was 444882, and 61 were followed-up. For the seven most significant meta-analysis p-values: the position (columns 1-3), the primary and follow-up study p-values (column 4 and 5), the meta-analysis p-values (column 6), and the r-values (column 7). See Table S2 of the SI for the results for all 61 SNPs followed-up. The lower bound for f_{00} was $l_{00} = 0.8$ for the r-value computation.

Chr.	Position	Gene	p1	p2	p_meta	r-value
6	32685358	HLA-DRB1	8.19e-08	8.57e-14	4.13e-20	0.0073
8	6810195	DEFAs	2.04e-07	1.25e-07	3.18e-14	0.009
6	32779226	HLA-DQA/B	3.28e-08	3.57e-06	3.43e-13	0.0058
22	28753460	MTMR3	2.30e-07	2.02e-05	1.17e-11	0.009
6	30049922	HLA-A	4.05e-09	3.68e-04	1.74e-11	0.009
17	7403693	TNFSF13	1.50e-06	2.52e-05	9.40e-11	0.038
17	7431901	MPDU1	5.52e-07	3.16e-04	4.31e-10	0.016

Our second example is GWAS of Crohn's disease (CD). To discover association between SNPs and CD, [15] examined 635547 SNPs on 3230 cases and 4829 controls of European descent, collected in three separate studies: NIDDK4, WTCCC5, and a Belgian-French study. For follow-up, 126 SNPs were measured in 2325 additional cases and 1809 controls as well as in an independent family-based dataset of 1339 trios of parents and their affected offspring. The two smallest p-values in each distinct region with primary study p-values below 5×10^{-5} were considered for follow-up. Table S3 in the SI shows the 126 SNPs followed-up. A replicability analysis with $l_{00} = 0.8$ identified 54 SNPs with r-values below 0.05. The 54 SNPs with replicated associations did not correspond to the 54 SNPs with smallest meta-analysis p-values. For example, the SNP in row 35 had the 35th smallest meta-analysis p-value, but its r-value was 0.09, thus it was not among the 54 replicated discoveries. In order to identify the number of SNPs with evidence of associations, the following typical meta-analysis can be done: applying the BH procedure at level 0.05 on the combined data from the primary and follow-up studies, by using the meta-analysis p-values when follow-up information was available and primary study p-values when there is no follow-up information. Since we only had primary and follow-up p-values for the 126 SNPs selected for follow-up, we set the remaining 635421 p-values conservatively as one. We discovered 76 SNPs associated with the disease in at least one of the studies, and the 54 replicated discoveries were a subset thereof. The last column of Table S3 in the SI shows the 30 SNPs that were highlighted as "convincingly (Bonferroni $P < 0.05$) replicated CD risk loci", based on the follow-up study p-values, in Table 2 of the main manuscript of [15]. These 30 SNPs have r-values below 0.05, so they are a subset of the 54 replicated discoveries. Our replicability analysis discovers more loci, in particular three loci (rows 34, 44, and 59 in Table S3 of the SI) that did not reach the conservative Bonferroni threshold of [15] on the follow-up study p-values, yet were pointed out in Table 3 of [15] to be "Nominally (uncorrected $P < 0.05$) replicated CD risk loci".

Our third example is GWAS of type 2 diabetes (T2D). To discover association between SNPs and T2D, [21] examined more than two million SNPs imputed from about 400000 SNPs collected on 4549 cases and 5579 controls combined from three separate studies: DGI, WTCCC, and FUSION. For follow-up, 68 SNPs were measured in 10037 cases and 12389 controls combined from additional genotyping of DGI, WTCCC, and FUSION. The 68 SNPs chosen for follow-up had primary study p-values below 10^{-4}, and they were in loci that were not discovered in previous studies. For additional follow-up, 11 out of the 68 SNPs were measured in 14157 cases and 43209 controls of European descent combined from 10 centers. The 11 SNPs forwarded for an additional follow-up had p-values below 0.005 in the first follow-up study, as well as a meta-analysis p-values below 10^{-5} when combining the evidence from the primary study and the first follow-up study. While there was no evidence of replicability from the primary study to the follow-up studies, there was evidence of replicability from the first follow-up study to the second follow-up study. Table 3 shows the 11 SNPs followed-up from the first follow-up study to the second follow-up study. Since most of the 68 SNPs in the first follow-up study are already believed to be associated with

the disease, the lower bound of $l_{00} = 0$, is the most appropriate. Five SNPs had r-values below 0.05.

Table 3: Replicability analysis for FDR control for the study of [21] on GWAS of T2D. The number of SNPs in the first follow-up study was 68, and 11 were followed-up to the second follow-up study. For these 11 SNPs: the positions (columns 1-2), the primary study p-values and first and second follow-up studies p-values (columns 3-5), the meta-analysis p-values from all 3 studies (column 6), and the r-values quantifying the evidence of replicability from the first to the second follow-up study. Note that the lower bound for f_{00} was $l_{00} = 0$ for the r-value computation, since the set of SNPs in the first follow-up study are already believed to be associated with T2D.

Chr.	Position	p.primary	p1	p2	p_meta	r-value
7	27953796	1.55e-04	8.07e-05	1.34e-07	4.96e-14	0.0055
10	12368016	4.21e-04	5.40e-05	1.49e-04	1.21e-10	0.0055
12	69949369	1.80e-05	9.83e-03	4.35e-05	1.11e-09	0.1490
2	43644474	1.83e-04	1.62e-03	9.22e-05	1.12e-09	0.0441
3	64686944	5.44e-04	1.02e-04	3.47e-03	1.17e-08	0.0254
1	120230001	1.14e-04	2.89e-03	1.95e-03	4.10e-08	0.0604
12	53385263	3.18e-05	3.11e-03	8.81e-03	1.79e-07	0.0604
3	12252845	1.05e-05	4.50e-03	1.22e-02	1.97e-07	0.0765
1	120149926	1.35e-03	1.17e-03	7.84e-03	4.04e-07	0.0431
6	43919740	5.41e-05	1.46e-03	9.49e-02	4.03e-06	0.2090
2	60581582	3.38e-05	1.38e-03	6.54e-01	1.02e-04	1.0000

1.3 Comparison with the procedure on maximum p-values

Another method for replicability analysis, which has been shown in [13] to be inferior to Procedure 1.1 with $l_{00} = 0$, is the following. For each hypothesis followed-up, the p-value for the eplicability analysis is the maximum of the two studies p-values. If the hypothesis is not followed-up, the p-value for the no replicability null hypothesis is one. The BH procedure is applied on the resulting m no replicablity null hypothesis p-values. This procedure can be viewed as applying steps 2-4 of Procedure 1.1 with parameters $(c_1, c_2) = (1, R_1/m)$, where R_1 is the number of hypotheses selected for follow-up. This alternative procedure is usually less powerful than the procedure with $c_2 = 0.5$, since 0.5 is far greater than R_1/m for GWAS on the whole genome. This comparison is not entirely fair, since the information on l_{00} can be incorporated to the BH procedure on the maximum p-values as well. The improved procedure can be viewed as applying steps 2-4 of Procedure 1.1 with parameters $(c_1, c_2) = (\frac{1}{1-l_{00}}, \frac{1}{1-l_{00}}\frac{R_1}{m})$, and with Step 1 of Procedure 1.1 modified to be $c_1 = \frac{1}{1-l_{00}}$. This procedure too will usually be less powerful than Procedure 1.1 with $c_2 = 0.5$ for $\frac{1}{1-l_{00}}\frac{R_1}{m} \ll 0.5$, as is typical for GWAS on the whole genome. If the follow-up study is extremely powerful, so that the p-values from non-null hypotheses in the follow-up

study are of order $\frac{1}{1-l_{00}}\frac{R_1}{m}$, then the BH procedure on maximum p-values may have a power advantage over Procedure 1.1. However, this is an uncommon setting for GWAS where only a tiny fraction of SNPs are examined in the follow-up study. Indeed, the alternative procedure performed worse than Procedure 1.1 in all the examples we considered. Specifically, with $q = 0.05$ and $l_{00} = 0.8$, the alternative procedure discovered in the first example only two SNPs compared to our seven discoveries, and in the second example only 22 SNPs compared to our 54 discoveries. In Table S1 we show the superior power of Procedure 1.1 over this alternative procedure in realistic GWAS simulations.

1.4 Choice of selection rule for replicability analysis

For a given FDR level q, the hypotheses that are promising to establish replicability using Procedure 1.1 are the set of hypotheses rejected with the BH procedure at level $c_1 q$ on the primary study p-values. Therefore, for the purpose of replicability analysis, the set of hypotheses to be considered should be only this set or a subset thereof. This means that if R_1 hypotheses are followed-up for meta-analysis, not all R_1 hypotheses need to be included in a replicability analysis at a predetermined level q. Specifically, in order for the r-value to be below q, only the subset of R_1 hypotheses selected for follow-up with primary study p-values that are small enough need to be considered, where our requirement for small enough is that the BH adjusted p-values of the R_1 primary study p-values, corrected for multiplicity of m hypotheses, are below $c_1 q$. Formally, if SNP $i \in \mathcal{R}_1$, it should be considered for replicability analysis only if $\min_{p_{1j} \geq p_{1i}, j \in \mathcal{R}_1} \frac{m p_{1j}}{rank(p_{1j})} \leq c_1 q$, where $rank(p_{1j})$ is the rank of p_{1j} among all primary study p-values (with maximum rank for ties). Computing the r-values for the subset of $\mathcal{R}_1$ with small enough primary study p-values, we receive smaller r-values than if all R_1 SNPs are considered for replicability analysis. For the first example, for an FDR level of 0.05, only 17 SNPs out of the 61 followed-up had primary study p-values small enough to be considered for replicability analysis. The number of r-values below 0.05 was still seven with this modified selection rule, but these seven r-values were smaller than the r-values for the seven SNPs in Table 2. Specifically, with parameters $(l_{00}, c_2) = (0.8, 0.5)$ for this superior selection rule that selected 17 SNPs for follow-up, the r-values were 0.005, 0.008, 0.005, 0.008, 0.005, 0.038, 0.016, whereas the r-values computed using all 61 SNPs selected were, respectively, 0.007, 0.009, 0.006, 0.009, 0.009, 0.038, and 0.016.

1.5 Procedure for FWER control

An FWER controlling procedure for replicability analysis was suggested in [13], that applies an FWER controlling procedure at level $c_1\alpha$ on the primary study p-values, and at level $c_2\alpha$ on the subset of discoveries from the primary study that were followed-

up, where $c_1 + c_2 = 1$. The FWER criterion,

$$FWER = \Pr(R_{00} + R_{01} + R_{10} > 0),$$

is more stringent than the FDR, yet it may be preferred over the FDR controlling procedure suggested above when very few hypotheses are followed-up. If a non-zero lower bound on f_{00} is available, then this lower bound can be used in order to choose parameters (c_1, c_2) with a sum greater than one. Specifically, for FWER control using Bonferroni, the adjusted p-values are

$$p_j^{Bonf-REPadj} = \max\left(mp_{1j}/c_1, |\mathcal{R}_1|p_{2j}/c_2\right),$$

and it is straightforward to show that the procedure that declares the hypotheses with adjusted p-values below α as replicated controls the FWER below the level

$$c_1 c_2 \alpha^2 f_{00} E\left(\frac{1}{R_1}\right) + f_{01} c_1 \alpha + c_2 \alpha E\left(\frac{|\mathcal{R}_1 \cap I_{10}|}{R_1}\right).$$

Therefore, for $l_{00} \leq f_{00}$, setting $c_1 = \frac{1-c_2}{1-l_{00}}$ will result in a FWER level below

$$\alpha + \frac{c_2(1-c_2)}{(1-l_{00})} l_{00} E(1/R_1)\alpha^2.$$

For a choice $c_2 = 0.5$, the second term in the upper bound is negligible for $l_{00} \leq 0.8$, and can thus be ignored. The discovery is considered replicated if the primary study p-value is below $[0.5/(1-l_{00})]\alpha/m$ and the follow-up p-value is below $0.5\alpha/R_1$.

As an example, consider the following GWAS of thyrotoxic periodic paralysis (TPP). To discover association between SNPs and thyrotoxic periodic paralysis, [17] measured 486782 SNPs in 70 cases and 800 controls from the Hong Kong (Southern) Chinese population. The 4 most significant SNPs were followed-up in an additional 54 southern Chinese TPP cases and 400 healthy Taiwanese controls. Table 4 shows the 4 SNPs followed-up. The associations were successfully replicated with adjusted p-values far below 0.05, concurring with the claim in [17] that "Associations for all four SNPs were successfully replicated".

Table 4: Replicability analysis for FWER control for the study of [17] on GWAS of TPP. The number of SNPs in the primary study was 486782, and four SNPs were followed-up. The lower bound for f_{00} was $l_{00} = 0.8$ for the r-value computation.

Chr.	Position	p1	p2	p_meta	r-value
17	65837933	6.28e-10	1.49e-05	7.69e-14	0.00012
17	65818432	1.39e-09	7.36e-05	1.59e-12	0.00059
17	65799923	2.27e-09	7.25e-05	1.09e-12	0.00058
17	65778654	1.84e-08	0.000116	1.6e-11	0.00358

2 Discussion

We proposed the r-value as an FDR-based measure of significance for replicability analysis. The method guarantees that the expected proportion of false replicability claims among the declared replicability claims remains small. We showed in examples that the smallest meta-analysis p-values may not have the strongest evidence towards replicability of association, and therefore we suggest to report the r-values in addition to the meta-analysis p-values in the table of results. Our web application http://www.math.tau.ac.il/∼ruheller/App.html accepts the primary and follow-up studies p-values of the hypotheses followed-up, as well as l_{00}, and computes the r-values.

In this work, we suggested r-values for FDR control as well as r-values for FWER control. The FDR may be preferred over the FWER for replicability analysis if it is enough to guarantee that the expected fraction of false replicability claims among the replicability claims is small. When only few hypotheses are followed-up, replicability analysis with FWER control may be applied.

We saw examples where the primary study was comprised of more than one study, as well as where more than one follow-up study was performed. In the current work, we use all the information from the primary studies for selection for follow-up, and in order to establish replicability the meta-analysis p-values of the primary studies and the meta-analysis p-values of the follow-up studies were used in Procedure 1.1. Alternative ways of combining the evidence, that can also point to the pair of studies in which the evidence of replication is strongest, will be considered in the future. The scientific evidence of two out of two $(2/2)$ studies is more convincing than that of two out of three $(2/3)$ studies or two out of n $(2/n)$ studies, and the scientific evidence of $3/n$ studies is more convincing than that of $2/n$ towards replicability. In the future, we plan to develop methods for computing the $r_{u/n}$-value, that quantifies the evidence that the finding has been replicated in at least u out of n studies, for $2 \leq u \leq n$. This problem has been addressed in [18], but as we showed in [13] alternatives along the lines of the procedures suggested here may benefit from increased power.

Finally, we would like to comment that our procedure should not be used when a single study is available and it is divided into two sub-studies, where one sub-study is set aside for validation. In this setting, the replicability analysis is meaningless since there are only two subsets of the same study instead of two independently conducted studies. In our notation, for the two sub-studies all hypotheses are necessarily in I_{00} or in I_{11}, since we only have two random samples from the same population. Therefore, discoveries from a typical meta-analysis are also discoveries from I_{11}. However, to detect associations there is no advantage in sub-dividing the data into two and use the first half to limit the features analyzed in the second half, since the power to detect associations will be lower than if all the data is analyzed together. A valid purpose for dividing the data in a single study into two sub-studies is the following. If the analysis is not thoroughly planned out before looking at the data, then the

first sub-study may serve for exploratory purposes as well as for planning the analysis of the second sub-study. Valid inference will then be obtained by the discoveries in the second sub-study only. By putting a subset aside for this purpose, the aim is to detect associations, and not to detect replicated associations. For detecting replicated associations, at least two independently conducted studies must be available.

References

[1] Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, 11: 241–247.

[2] McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J., and Hirschhorn, J. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews Genetics*, pages 356–369.

[3] NCI-NHGRI (2007). Replicating genotype-phenotype associations. *Nature*, 447(7):655–660.

[4] Kraft, P., Zeggini, E., and Ioannidis, J. (2009). Replication in genome-wide association studies. *Statistical science*, 24 (4):561–573.

[5] Yu et al. (2012). A genome-wide association study in Han Chinese identifies multiple susceptibility loci for IgA nephropathy. *Nature Genetics*, 44:178–182.

[6] Peng, R. (2009). Reproducible research and Biostatistics *Biostatistics*, 10 (3): 405-408

[7] Storey, J. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64 (3):479–498.

[8] Westfall,P.H. and Young, S.S.(1989). P-value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association*, 84:780–786.

[9] Li, Q. and and Brown, J. and Huang, H. and Bickel, P. (2011). Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5 (3):1752–1779.

[10] Natarajan, L., Pu, M., and Messer, K. (2012). Statistical tests for the intersection of independent lists of genes: sensitivity, FDR, and type I error control. *The Annals of Applied Statistics*, 6 (2):521-541.

[11] Heller, R. and Yekutieli, D.(2013). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics*, conditionally accepted.

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

[13] Bogomolov, M. and Heller, R. (2013). Discovering findings that replicate from a primary study of high dimension to a follow-up study. *Journal of the American Statistical Association*, accepted.

[14] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, 57 (1):289–300.

[15] Barret et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics*, 40: 955–962.

[16] Zeggini, E., Scott, L., Saxena, R., Voight, and DIAGRAM (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, 40:638–645.

[17] Cheung et al. (2012). Genome-wide association study identifies a susceptibility locus for thyrotoxic periodic paralysis at 17q24.3. *Nature Genetics*, 44:1026–1029.

[18] Benjamini, Y. and Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics*, 64:1215–1222.

[19] Su, Z., Marchini, J., and Donnelly, P. (2011). Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27 (16):2304–2305.

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

[21] Zeggini, E., Scott, L., Saxena, R., Voight, and DIAGRAM (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics*, 40:638–645.

A The upper bound on the FDR of Procedure 1

The rule by which the set $\mathcal{R}_1$ is selected needs to be a stable selection rule.

Definition A.1 *[13] A* stable selection rule *for Procedure 1 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 j is still selected, will not change the set $\mathcal{R}_1$.*

This is a technical condition that is typically satisfied. Stable selection rules include selecting the hypotheses with p-values below a certain cut-off, or by a non-adaptive multiple testing procedure such as the BH procedure for FDR control on the primary study, or selecting the k hypotheses with the smallest p-values, where k is fixed in advance.

Theorem A.1 *Procedure 1 controls the FDR at level*

$$f_{00}c_1c_2q^2 + f_{01}c_1q + E\left(\frac{|I_{10} \cap \mathcal{R}_1|}{|\mathcal{R}_1|}\right) c_2q$$

if the rule by which the set $\mathcal{R}_1$ is selected is a stable selection rule, the p-values within the follow-up study are jointly independent or have property PRDS, and are independent of the primary study p-values, in either one of the following situations:

1. *The p-values within the primary study are jointly independent.*
2. *Arbitrary dependence among the p-values within the primary study, and the primary study p-values of the hypotheses that are selected for follow-up are at most $\frac{c_1q}{m}$.*

Proof. We recall the following definitions from [13]. 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 fixed as long as P_{1j} is such that j is selected based on $(P_1^{(j)}, P_{1j})$. For any $j \in \{1, \ldots, m\}$ and given $P_1^{(j)}$, for $i \in \{1, \ldots, j-1, j+1, \ldots, m\}$

$$T_i = \begin{cases} \max\left(\frac{mP_{1i}}{c_1q}, \frac{(|\mathcal{R}_1^{(j)}(P_1^{(j)})|+1)P_{2i}}{c_2q}\right) & \text{if } i \in \mathcal{R}_1^{(j)}(P_1^{(j)}), \\ \infty & \text{otherwise.} \end{cases}$$

Let $T_{(1)} \leq \ldots \leq T_{(m-1)}$ 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 $j \in I_{00} \cup I_{01} \cup I_{10}$ is declared replicated by Procedure 1, r hypotheses are declared replicated including j:

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

Note that given P_1, for $r > |\mathcal{R}_1|$, $C_r^{(j)} = \emptyset$, since exactly $|\mathcal{R}_1|$ 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 $(P_1^{(j)}, P_2^{(j)})$. Therefore,

$$\sum_{r=1}^{m} \Pr\left(C_r^{(j)}\right) = 1. \tag{1}$$

Let R_j be the indicator of whether j was declared replicated for $j = 1, \ldots, m$, and $R = \sum_{j=1}^m R_j$. The FDR for replicability analysis is

$$FDR = \sum_{x \in \{00,01,10\}} \sum_{j \in I_x} E\left(\frac{R_j}{\max(R, 1)}\right).$$

For Procedure 1,

$$E\left(\frac{R_j}{\max(R,1)}\right) = \sum_{r=1}^{m} \frac{1}{r}\Pr\left(j \in \mathcal{R}_1, P_{1j} \leq \frac{rc_1q}{m}, P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1|}, C_r^{(j)}\right). \tag{2}$$

For the proof of each one of the items, we will give an upper bound for expression (2) for $j \in I_{01}$, $j \in I_{10}$, and $j \in I_{00}$. We will start with the proof of item 1 for the case where the p-values within each study are jointly independent. This proof is similar to the proof given in [13], where the upper bound on the FDR of Procedure 1 was derived for $c_1q = c_2q = q'$. For completeness, the proof for this case is given below. For $j \in I_{01}$,

$$\begin{aligned}
&\sum_{r=1}^{m} \frac{1}{r}\Pr\left(j \in \mathcal{R}_1, P_{1j} \leq \frac{rc_1q}{m}, P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1|}, C_r^{(j)}\right) \\
&\leq \sum_{r=1}^{m} \frac{1}{r}\Pr\left(P_{1j} \leq \frac{rc_1q}{m}, C_r^{(j)}\right) \\
&\leq \frac{c_1q}{m}\sum_{r=1}^{m}\Pr\left(C_r^{(j)}\right) &(3)\\
&= \frac{c_1q}{m}, &(4)
\end{aligned}$$

where inequality (3) follows from the independence of the p-values and the fact that P_{1j} is a null-hypothesis p-value, and equality (4) follows from (1). For $j \in I_{00}$, (2) is smaller or equal to

$$\begin{aligned}
&\sum_{r=1}^{m} \frac{1}{r}\Pr\left(P_{1j} \leq \frac{rc_1q}{m}, P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1|}, C_r^{(j)}\right) \\
&\leq \sum_{r=1}^{m} \frac{1}{r}\Pr\left(P_{1j} \leq \frac{rc_1q}{m}, P_{2j} \leq \frac{|\mathcal{R}_1|c_2q}{|\mathcal{R}_1|}, C_r^{(j)}\right) &(5)\\
&\leq c_2q\frac{c_1q}{m}\sum_{r=1}^{m}\Pr(C_r^{(j)}) &(6)\\
&= c_2q\frac{c_1q}{m}, &(7)
\end{aligned}$$

where inequality (5) follows from the fact that for any given realization of $|\mathcal{R}_1|$ and value of r such that $r > |\mathcal{R}_1|$, $C_r^{(j)} = \emptyset$, inequality (6) follows from the independence of the p-values and the fact that P_{1j} and P_{2j} are null-hypothesis p-values, and equality

(7) follows from (1). For $j \in I_{10}$ and an arbitrary fixed $p_1 = (p_{11}, \ldots, p_{1m})$

$$
\begin{aligned}
&E\left(\frac{R_j}{\max(R,1)}|P_1 = p_1\right) = \\
&\sum_{r=1}^{|\mathcal{R}_1(p_1)|} \frac{1}{r}\Pr\left(j \in \mathcal{R}_1(p_1), P_{1j} \leq \frac{rc_1q}{m}, P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1(p_1)|}, C_r^{(j)}|P_1 = p_1\right) \\
&\leq I\left(j \in \mathcal{R}_1(p_1)\right) \sum_{r=1}^{|\mathcal{R}_1(p_1)|} \frac{1}{r}\Pr\left(P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1(p_1)|}, C_r^{(j)}|P_1 = p_1\right) \\
&= I\left(j \in \mathcal{R}_1(p_1)\right) \sum_{r=1}^{|\mathcal{R}_1(p_1)|} \frac{1}{r}\Pr\left(C_r^{(j)}|P_1 = p_1\right) \qquad (8) \\
&\qquad \times\Pr\left(P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1(p_1)|}|P_1 = p_1\right) \\
&\leq \frac{c_2q}{|\mathcal{R}_1(p_1)|} I\left(j \in \mathcal{R}_1(p_1)\right) \sum_{r=1}^{|\mathcal{R}_1(p_1)|} \Pr\left(C_r^{(j)}|P_1 = p_1\right) \qquad (9) \\
&= \frac{c_2q}{|\mathcal{R}_1(p_1)|} I\left(j \in \mathcal{R}_1(p_1)\right), \qquad (10)
\end{aligned}
$$

where equality (8) follows from the conditional independence of $C_r^{(j)}$ and the event $\{P_{2j} \leq rc_2q/|\mathcal{R}_1(p_1)|\}$, inequality (9) follows from the independence of the p-values and the fact that P_{2j} is a null-hypothesis p-value, and equality (10) 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 that for $j \in I_{10}$

$$
E\left(\frac{R_j}{\max(R,1)}\right) \leq c_2qE\left(\frac{I\left(j \in \mathcal{R}_1\right)}{|\mathcal{R}_1|}\right). \qquad (11)
$$

The FDR for replicability analysis is therefore bounded by

$$
\begin{aligned}
FDR = &\sum_{x \in \{00,01,10\}} \sum_{j \in I_x} E\left(\frac{R_j}{\max(R,1)}\right) \leq \\
&f_{00}c_1c_2q^2 + f_{01}c_1q + c_2qE\left(\frac{|I_{10} \cap \mathcal{R}_1|}{|\mathcal{R}_1|}\right),
\end{aligned}
$$

where the inequality follows from summing over the upper bounds (4), (7), and (11), thus completing the proof of item 1 for the case where the p-values are independent within each study.

We will now prove item 1 for the case where the p-values within the primary study are jointly independent and the p-values within the follow-up study have property PRDS. Note that the upper bound in (4) is derived only using the independence of the p-values within the primary study and independence of the p-values across the

studies, therefore this upper bound holds when the set of p-values within the follow-up study has property PRDS. Hence, for $j \in I_{01}$ an upper bound for expression (2) is $c_1 q/m$. For $j \in I_{00} \cup I_{10}$ and an arbitrary fixed $p_1 = (p_{11}, \ldots, p_{1m})$

$$
\begin{aligned}
&E\left(\frac{R_j}{\max(R,1)}|P_1 = p_1\right) = \\
&\sum_{r=1}^{|\mathcal{R}_1(p_1)|} \frac{1}{r}\Pr\left(j \in \mathcal{R}_1(p_1), P_{1j} \leq \frac{rc_1q}{m}, P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1(p_1)|}, C_r^{(j)}|P_1 = p_1\right) \\
&\leq I\left(p_{1j} \leq \frac{|\mathcal{R}_1(p_1)|c_1q}{m}, j \in \mathcal{R}_1(p_1)\right) \sum_{r=1}^{|\mathcal{R}_1(p_1)|} \frac{1}{r}\Pr\left(P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1(p_1)|}, C_r^{(j)}|P_1 = p_1\right) \\
&= I\left(p_{1j} \leq \frac{|\mathcal{R}_1(p_1)|c_1q}{m}, j \in \mathcal{R}_1(p_1)\right) \sum_{r=1}^{|\mathcal{R}_1(p_1)|} \frac{1}{r}\Pr\left(C_r^{(j)}|P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1(p_1)|}, P_1 = p_1\right) \\
&\qquad \times \Pr\left(P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1(p_1)|}|P_1 = p_1\right) \\
&\leq \frac{c_2q}{|\mathcal{R}_1(p_1)|} I\left(p_{1j} \leq \frac{|\mathcal{R}_1(p_1)|c_1q}{m}, j \in \mathcal{R}_1(p_1)\right) \qquad (12) \\
&\qquad \times \sum_{r=1}^{|\mathcal{R}_1(p_1)|} \Pr\left(C_r^{(j)}|P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1(p_1)|}, P_1 = p_1\right) \\
&\leq \frac{c_2q}{|\mathcal{R}_1(p_1)|} I\left(p_{1j} \leq \frac{|\mathcal{R}_1(p_1)|c_1q}{m}, j \in \mathcal{R}_1(p_1)\right), \qquad (13)
\end{aligned}
$$

where inequality (12) follows from the independence of the p-values across the studies and the fact that P_{2j} is a null-hypothesis p-value. We will now show that inequality (13) holds. It follows from item 1 of Lemma S2.1 in the Supplementary Material of [13] that

$$
\sum_{r=1}^{|\mathcal{R}_1(p_1)|} \Pr\left(C_r^{(j)}|P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1(p_1)|}, P_1 = p_1\right) \leq 1
$$

for $j \in I_{10} \cap \mathcal{R}_1(p_1)$. It is straightforward to verify that this result holds for $j \in I_{00} \cap \mathcal{R}_1(p_1)$ as well, yielding inequality (13). Using (13) we obtain the upper bounds on expression (2) for $j \in I_{00}$ and for $j \in I_{10}$. For $j \in I_{10}$, it follows that

$$
E\left(\frac{R_j}{\max(R,1)}|P_1 = p_1\right) \leq \frac{c_2q}{|\mathcal{R}_1(p_1)|} I\left(j \in \mathcal{R}_1(p_1)\right),
$$

therefore

$$
E\left(\frac{R_j}{\max(R,1)}\right) \leq c_2 q E\left(\frac{I\left(j \in \mathcal{R}_1\right)}{|\mathcal{R}_1|}\right). \qquad (14)
$$

For $j \in I_{00}$, it follows that

$$
E\left(\frac{R_j}{\max(R,1)}\right) \leq c_2 q E\left[\frac{I\left(P_{1j} \leq \frac{|\mathcal{R}_1(P_1)|c_1q}{m}, j \in \mathcal{R}_1(P_1)\right)}{|\mathcal{R}_1(P_1)|}\right] \qquad (15)
$$

Note that for $j \in I_{00}$

$$
\begin{aligned}
E\left[\frac{I\left(P_{1j} \leq \frac{|\mathcal{R}_1(P_1)|c_1q}{m}, j \in \mathcal{R}_1(P_1)\right)}{|\mathcal{R}_1(P_1)|}\right] &= \sum_{r=1}^{m} \frac{1}{r}\Pr\left(P_{1j} \leq \frac{rc_1q}{m}, j \in \mathcal{R}_1(P_1), |\mathcal{R}_1^{(j)}(P_1^{(j)})| = r\right) \\
&\leq \sum_{r=1}^{m} \frac{1}{r}\Pr\left(P_{1j} \leq \frac{rc_1q}{m}, |\mathcal{R}_1^{(j)}(P_1^{(j)})| = r\right) \\
&\leq \frac{c_1q}{m}\sum_{r=1}^{m}\Pr\left(|\mathcal{R}_1^{(j)}(P_1^{(j)})| = r\right) = \frac{c_1q}{m}. \qquad (16)
\end{aligned}
$$

The inequality in (16) follows from the independence of the p-values within the primary study and the fact that P_{1j} is a null-hypothesis p-value. The equality in (16) follows from the fact that $\cup_{r=1}^{m}\{|R_1^{(j)}(P_1^{(j)})| = r\}$ is the entire space of $P_1^{(j)}$, represented as a union of disjoint events. Combining (15) with (16) we obtain for $j \in I_{00}$

$$
E\left(\frac{R_j}{\max(R,1)}\right) \leq c_2q\frac{c_1q}{m}. \qquad (17)
$$

The FDR for replicability analysis is therefore bounded by

$$
\begin{aligned}
FDR = \sum_{x\in\{00,01,10\}}\sum_{j\in I_x} E\left(\frac{R_j}{\max(R,1)}\right) \leq \\
f_{00}c_1c_2q^2 + f_{01}c_1q + c_2qE\left(\frac{|I_{10}\cap\mathcal{R}_1|}{|\mathcal{R}_1|}\right),
\end{aligned}
$$

where the inequality follows from summing over the upper bounds (4), (14), and (17), thus completing the proof of item 1 for the case where the p-values within the primary study are independent and the set of p-values within the follow-up study has property PRDS.

For item 2, we note that when one of the selection criteria for follow-up is that the primary study p-values are below $\frac{c_1q}{m}$, we have that for $r \in \{1,\ldots,m\}$,

$$
\left\{P_1 :\ j \in \mathcal{R}_1(P_1), P_{1j} \leq \frac{rc_1q}{m}\right\} \subseteq \left\{P_1 :\ P_{1j} \leq \frac{c_1q}{m}\right\}. \qquad (18)
$$

Therefore, for $j \in I_{01}$,

$$
\begin{aligned}
E\left(\frac{R_j}{\max(R,1)}\right) &\leq \sum_{r=1}^{m}\frac{1}{r}\Pr\left(P_{1j} \leq \frac{c_1q}{m}, P_{2j} \leq \frac{rc_2q}{|\mathcal{R}_1|}, C_r^{(j)}\right) \\
&\leq \sum_{r=1}^{m}\frac{1}{r}\Pr\left(C_r^{(j)}|P_{1j} \leq \frac{c_1q}{m}\right)\Pr\left(P_{1j} \leq \frac{c_1q}{m}\right) \\
&\leq \frac{c_1q}{m}\sum_{r=1}^{m}\Pr\left(C_r^{(j)}|P_{1j} \leq \frac{c_1q}{m}\right) \qquad (19) \\
&= \frac{c_1q}{m}, \qquad (20)
\end{aligned}
$$

where inequality (19) follows from the fact that P_{1j} is a null-hypothesis p-value, and equality (20) follows from the fact that $\cup_{r=1}^{m} C_r^{(j)}$ is the entire space, represented as a union of disjoint events.

For the case where the p-values within the follow-up study are independent, the derivation of the upper bounds (7) and (11) is the same as in item 1. The result follows from summing over the upper bounds (7), (11), and (20), as in the end of the proof of item 1. For the case where the p-values within the follow-up study have property PRDS, the derivation of the upper bound for $j \in I_{10}$ in (14) is the same as in item 1. The derivation of the upper bound for $j \in I_{00}$ is less involved than in item 1. Using (13) and (18) we obtain for $j \in I_{00}$

$$E\left(\frac{R_j}{\max(R,1)}|P_1 = p_1\right) \leq c_2 q I\left(p_{1j} \leq \frac{c_1 q}{m}\right),$$

yielding that

$$E\left(\frac{R_j}{\max(R,1)}\right) \leq c_2 q \frac{c_1 q}{m}, \tag{21}$$

since P_{1j} is a null-hypothesis p-value. The result follows from summing over the upper bounds (14), (20), and (21).

B Power comparison for different values of (l_{00}, c_2)

We conducted simulations in order to investigate how the power and FDR of Procedure 1 depend on its parameters (l_{00}, c_2), where $c_2 \in (0,1)$, for $l_{00} \in \{0.5, 0.8, 0.9\}$. In all simulations, in Procedure 1 step 1 we used the approximation $c_1 \approx \frac{1-c_2}{1-l_{00}}$, so there was no need to prefix q for the r-value computation. The p-values were generated independently as follows. Let H_j, $j = 1, \ldots, m$, be the hypotheses examined in the primary study, and P_{1j} and P_{2j} be the p-values for H_j in the primary and in the follow-up study respectively. We set $P_{1j} = 1 - \Phi(X_{1j})$ and $P_{2j} = 1 - \Phi(X_{2j})$, where $X_{1j} \sim N(\mu_{1j}, 1)$, $X_{2j} \sim N(\mu_{2j}, 1)$. For $i \in \{1,2\}$, we set $\mu_{ij} = 0$ if H_j is a true null hypothesis in study i, and $\mu_{ij} = \mu_i$ if H_j is a false null hypothesis in study i . The values of μ_1 and μ_2 were set according to the requirement that the power of Bonferroni procedure at level 0.05 in the primary study is π_1, and in the follow-up study is π_2, for $\pi_1 = 0.1$ and $\pi_2 \in \{0.2, 0.5, 0.8\}$. Specifically, we set $\mu_1 = \Phi^{-1}(1 - 0.05/m) - \Phi^{-1}(1 - \pi_1)$, and $\mu_2 = \Phi^{-1}(1 - 0.05/R_1) - \Phi^{-1}(1 - \pi_2)$, where Φ^{-1} is the inverse of the cumulative distribution function of a standard normal variable and R_1 is the number of rejected hypotheses by the BH procedure at level $c_1 \times 0.05$ applied on the primary study p-values. The total number of hypotheses, m, was set to be 1000, and $f_{ij} = |I_{ij}|/m$ for $i, j \in \{0,1\}$ were as follows: $f_{00} = 0.9$, $f_{01} = f_{10} = 0.025$, $f_{11} = 0.05$.

The simulation results were based on 1000 repetitions. The FDR was estimated by averaging the false discovery proportion. The average power was estimated by the

average number of true replicability claims, divided by mf_{11}. We also estimated the probability that Procedure 1 makes at least one true replicability claim (which we refer to as "power for at least one") by the proportion of repetitions in which at least one true replicability claim was made. The standard errors of the estimators of FDR and average power were of the order of 10^{-3}, while the standard error of the estimator of power for at least one was of the order of 10^{-2} for some sets of parameters, and was always below 0.016.

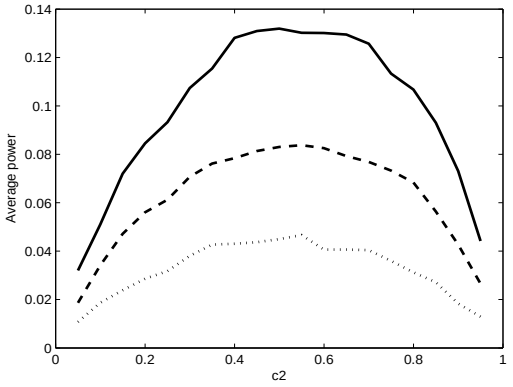
Figure S1 shows the average power and the power for at least one of Procedure 1 as a function of $c_2 \in \{0.05, 0.1, \ldots, 0.95\}$, for $l_{00} \in \{0.5, 0.8, 0.9\}$. As expected, both measures of power increase as l_{00} becomes higher. For fixed l_{00} and (π_1, π_2), the highest average power among all the choices of c_2 is close to the average power when $c_2 = 0.5$ (Figure S1, left column). This is also shown in Table 2 of the main manuscript. From the graphs of power for at least one as a function of c_2 it can be also seen that the choice $c_2 = 0.5$ is very close to optimal, and there is little sensitivity to the choice of c_2 as long as it is not too far from 0.5 (Figure S1, right column).

Figure S2 shows the FDR of Procedure 1 as a function of $c_2 \in \{0.05, 0.1, \ldots, 0.95\}$, for $l_{00} \in \{0.5, 0.8, 0.9\}$. It can be seen that the FDR is far below 0.05 for all the sets of parameters considered. This follows from the fact that our data generation may result in FDR much lower than the upper bound given in expression [1] of the main manuscript. In order to see this, note that it follows from the proof of Theorem A.1 that the FDR of Procedure 1 achieves the upper bound in expression [1] of the main manuscript when the p-values under the alternative are practically zero. In our simulation setting, this condition would hold if μ_i, for $i \in \{1, 2\}$ were always extremely large when compared to $N(0, 1)$ random variables, e.g. $\mu_i \geq 4$. Obviously this does not hold for our data generation process. Therefore we could get higher FDR values for another data generation process, however we still would not expect to achieve 0.05 because of using conservative upper bounds for f_{01} and $E(|I_{10} \cap \mathcal{R}_1|/|\mathcal{R}_1|)$ in expresson [1] of the main manuscript.

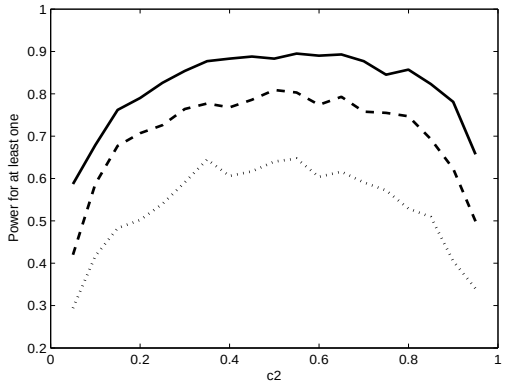
C GWAS simulation example

The goal of the simulation was threefold. First, to verify that the FDR is controlled below the nominal level for realistic simulations with GWAS type dependency, even if hypotheses with primary study p-values above $c1q/m$ are followed-up. Second, to compare the performance of the suggested Procedure 1 with the BH procedure on maximum p-values. Third, to examine the effect of l_{00} on the power of the two procedures.

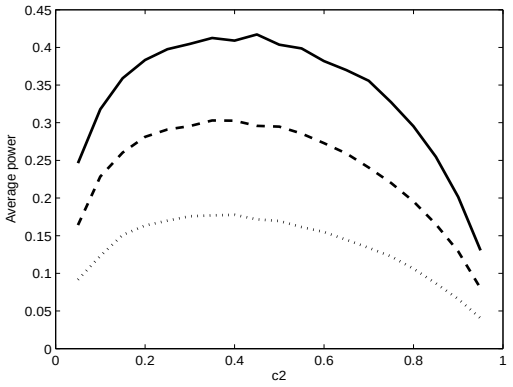
We simulated two GWAS from the simulator HAPGEN2 [19]. The two studies were generated from two samples of the HapMap project [20], a sample of 165 Utah residents with Northern and Western European ancestry (CEU), and a sample of 109 Chinese in Metropolitan Denver, Colorado (CHD). In the CEU and CHD populations,



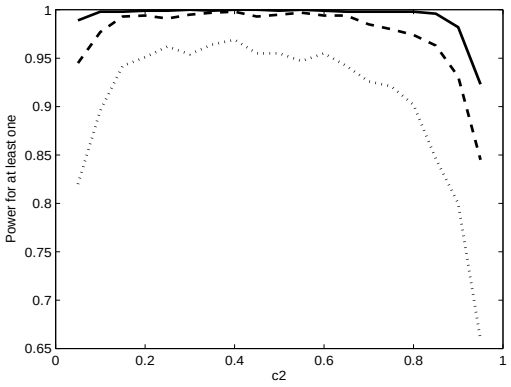
(a) Average power, $(\pi_1, \pi_2) = (0.1, 0.2)$.



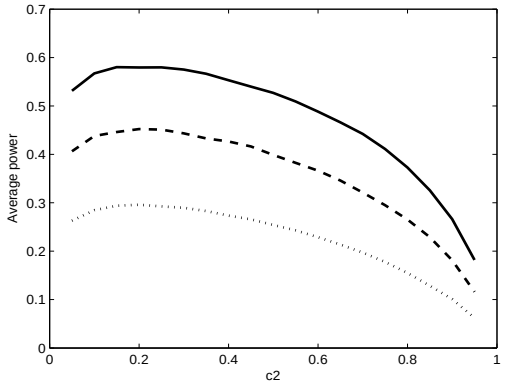
(b) Power for at least one, $(\pi_1, \pi_2) = (0.1, 0.2)$.



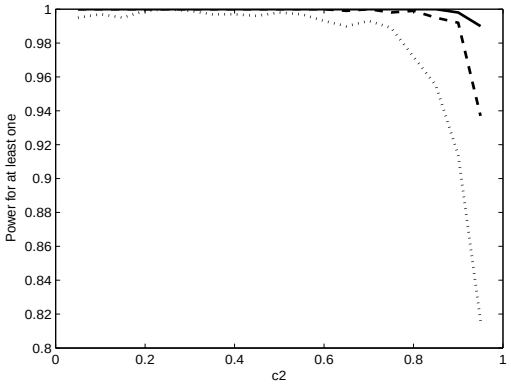
(c) Average power, $(\pi_1, \pi_2) = (0.1, 0.5)$.



(d) Power for at least one, $(\pi_1, \pi_2) = (0.1, 0.5)$.

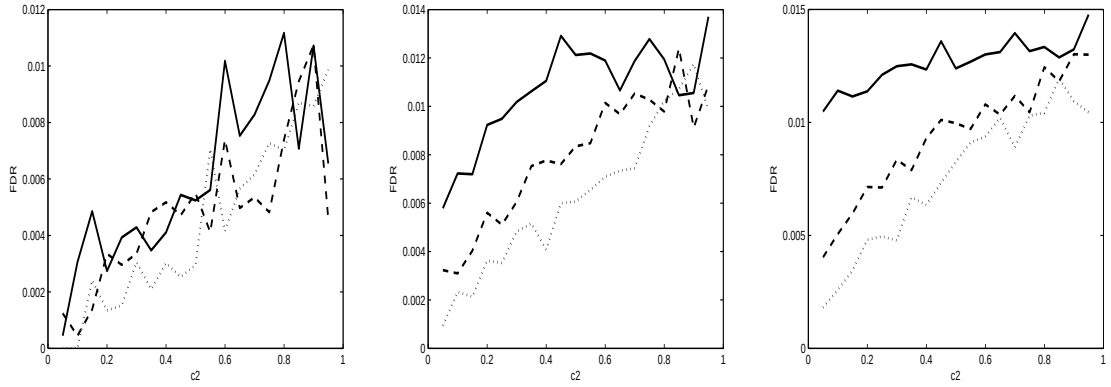


(e) Average power, $(\pi_1, \pi_2) = (0.1, 0.8)$.



(f) Power for at least one, $(\pi_1, \pi_2) = (0.1, 0.8)$.

Figure S1: The estimated average power (first column) and the probability of at least one true replicability claim (power for at least one, column 2) of Procedure 1 with parameters $(l_{00}, c_2, 0.05)$ as a function of c_2 in a simulation where $f_{00} = 0.9, f_{01} = f_{10} = 0.025, f_{11} = 0.05$, the number of hypotheses examined in the primary study is 1000, and the signal to noise ratios for the primary study and the follow-up study, μ_1/σ_1 and μ_2/σ_2, respectively, are taken according to the requirement that the power of Bonferroni procedure at level 0.05 in the primary study is π_1, and in the follow-up study is π_2 for $(\pi_1, \pi_2) = (0.1, 0.2)$ (row 1), $(\pi_1, \pi_2) = (0.1, 0.5)$ (row 2), $(\pi_1, \pi_2) = (0.1, 0.8)$ (row 3); $l_{00} = 0.9$ (solid), $l_{00} = 0.8$ (dashed), and $l_{00} = 0.5$ (dotted). The standard errors were below 0.016 for the estimator of power for at least one, and were of the order of 10^{-3} for the estimator of average power.



(a) $\pi_1 = 0.1$, $\pi_2 = 0.2$ (b) $\pi_1 = 0.1$, $\pi_2 = 0.5$ (c) $\pi_1 = 0.1, \pi_2 = 0.8$

Figure S2: The estimated FDR of Procedure 1 with parameters $(l_{00}, c_2, 0.05)$ as a function of c_2 in a simulation where $f_{00} = 0.9, f_{01} = f_{10} = 0.025, f_{11} = 0.05$, the number of hypotheses examined in the primary study is 1000, and the signal to noise ratios for the primary study and the follow-up study, μ_1/σ_1 and μ_2/σ_2, respectively, are taken according to the requirement that the power of Bonferroni procedure at level 0.05 in the primary study is π_1, and in the follow-up study is π_2 for $(\pi_1, \pi_2) = (0.1, 0.2)$ (left panel), $(\pi_1, \pi_2) = (0.1, 0.5)$ (middle panel), $(\pi_1, \pi_2) = (0.1, 0.8)$ (right panel); $l_{00} = 0.9$ (solid), $l_{00} = 0.8$ (dashed), and $l_{00} = 0.5$ (dotted). The standard errors were of the order of 10^{-3} for all the sets of parameters.

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. See [13] for the details of this simulation.

The CHD study was the primary study, and the CEU study was the follow-up study. Hypotheses were selected for follow-up only if they were discovered by the BH procedure at level $c_1 q$. If more than 300 hypotheses were discovered, then only the 300 hypotheses with smallest primary study p-values were selected for follow-up. Table S1 presents the average number of replicated findings, as well as the average FDP, for different values of l_{00}. From columns 4 and 7 it is clear that the FDR is controlled and that Procedure 1 is actually conservative, for all values of l_{00}. From a comparison of columns 2 and 5 it is clear that Procedure 1 is more powerful than the BH procedure on maximum p-values. Finally, from comparisons of the rows it is clear that the power increases as l_{00} increases.

Table S1: 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 values of l_{00}. The actual value of f_{00} was above 0.999. Results are given for Procedure 1 with parameters $(l_{00}, c_2) = (l_{00}, 0.5)$ in columns 2-4, and for the BH procedure on maximum p-values in columns 5-7. In Procedure 1 step 1 we used the approximation $c_1 \approx \frac{1-c_2}{1-l_{00}}$, so there was no need to prefix q for the r-value computation. The CHD study was the primary study, and the CEU study was the follow-up study. Hypotheses were selected for follow-up only if they were discovered by the BH procedure at level $c_1 \times 0.05$. If more than 300 hypotheses were discovered, then only the 300 hypotheses with smallest primary study p-values were selected for follow-up.

	Procedure 1 with parameters $(l_{00}, 0.5, 0.05)$			BH Procedure on maximum p-values		
	# Replicated findings		FDP	# Replicated findings		FDP
l_{00}	associated SNPs (SE)	disease SNPs (SE)	(SE)	associated SNPs (SE)	disease SNPs (SE)	(SE)
0	41.5 (5.3)	8.3 (0.5)	0.018 (0.015)	29.2 (3.2)	7.4 (0.4)	0.000 (0.000)
0.8	56.3 (5.2)	9.3 (0.4)	0.018 (0.016)	40.4 (3.8)	8.8 (0.4)	0.000 (0.000)
0.9	61.7 (4.9)	10.1 (0.4)	0.016 (0.014)	48.1 (4.0)	9.6 (0.4)	0.000 (0.000)
0.95	67.2 (5.0)	10.4 (0.4)	0.015 (0.013)	54.8 (4.1)	10.3 (0.3)	0.000 (0.000)
0.99	71.5 (4.2)	10.7 (0.2)	0.015 (0.013)	61.7 (3.8)	10.4 (0.3)	0.001 (0.001)

D GWAS Real data examples

Tables 1 and 2 show the results of the replicability analysis for the SNPs followed-up based on the results of the primary study (or studies). Columns 1-3 contain the position of each SNP. Columns 4-5 show the primary and follow-up p-values. Columns 6-8 show the r-values for different choices of (l_{00}, c_2). Column 9 shows the meta-analysis p-values, which are the unadjusted p-values computed using the data from the primary and follow-up studies for testing the global null hypothesis of no association in any of the studies. The rows are sorted by the meta-analysis p-values. The handful of findings with most significant meta-analysis p-values which were reported as interesting in the published works are marked with an $*$ in the last column.

Table S2: Replicability analysis for the study of [5]: GWAS of IgA nephropathy in Han Chinese. The number of SNPs in the primary study was 444882, and 61 were followed-up.

Chr.	Position	Gene	p1	p2	$(l_{00}, c_2) = (0, 0.5)$	$(l_{00}, c_2) = (0.5, 0.5)$	$(l_{00}, c_2) = (0.8, 0.5)$	p_meta	
6	HLA-DRB1	32685358	8.19e-08	8.57e-14	0.024	0.015	0.0073	4.13e-20	*
8	DEFAs	6810195	2.04e-07	1.25e-07	0.041	0.02	0.009	3.18e-14	*
6	HLA-DQA/B	32779226	3.28e-08	3.57e-06	0.022	0.015	0.0058	3.43e-13	*
22	MTMR3	28753460	2.30e-07	2.02e-05	0.041	0.02	0.009	1.17e-11	*
6	HLA-A	30049922	4.05e-09	3.68e-04	0.022	0.015	0.009	1.74e-11	*
17	TNFSF13	7403693	1.50e-06	2.52e-05	0.19	0.095	0.038	9.40e-11	*
17	MPDU1	7431901	5.52e-07	3.16e-04	0.082	0.041	0.016	4.31e-10	*
2	ACOXL	111315937	6.83e-05	3.41e-03	1	1	1	4.08e-07	
16	x	31255249	6.67e-05	7.41e-03	1	1	1	4.64e-06	
4	x	78121177	3.14e-10	8.16e-01	1	1	1	2.23e-05	
11	x	113369319	1.82e-09	9.74e-01	1	1	1	5.42e-05	
7	BBS9	33386800	2.75e-05	1.67e-01	1	1	1	1.17e-04	
11	x	44042263	1.74e-05	2.72e-01	1	1	1	1.24e-04	
4	x	40144579	9.95e-07	6.72e-01	1	1	1	1.85e-04	
12	x	13229380	1.23e-05	4.41e-01	1	1	1	3.09e-04	
14	x	69116920	4.60e-05	3.72e-01	1	1	1	3.71e-04	
8	x	30305114	3.19e-05	4.73e-01	1	1	1	5.38e-04	
12	x	129587780	4.59e-05	5.53e-01	1	1	1	6.84e-04	
6	x	31382359	8.20e-08	9.53e-01	1	1	1	7.64e-04	
16	WWOX	77632003	7.20e-05	4.57e-01	1	1	1	1.04e-03	
8	PTDSS1	97393458	5.67e-05	6.12e-01	1	1	1	1.09e-03	
6	x	26384629	4.32e-06	2.79e-01	1	1	1	1.25e-03	
13	x	62434248	3.77e-05	5.39e-01	1	1	1	1.70e-03	
11	FDX1	109836841	7.15e-05	7.19e-01	1	1	1	2.03e-03	
18	x	35923102	4.35e-05	2.85e-01	1	1	1	2.32e-03	
6	RANBP9	13733392	1.70e-05	8.45e-01	1	1	1	2.55e-03	
9	PSAT1	78162069	5.98e-05	8.01e-01	1	1	1	2.71e-03	
10	x	55006847	7.93e-05	6.87e-01	1	1	1	3.16e-03	
6	x	33163516	1.46e-04	7.74e-01	1	1	1	4.56e-03	
7	x	158006056	9.26e-05	7.50e-01	1	1	1	4.99e-03	
6	x	106231017	6.19e-05	8.59e-01	1	1	1	6.41e-03	
21	x	19339830	7.81e-05	5.34e-01	1	1	1	6.58e-03	
12	AEBP2	19488937	4.95e-05	4.77e-01	1	1	1	6.92e-03	
18	x	57221085	8.62e-06	5.48e-01	1	1	1	7.96e-03	
10	DUSP13	76538473	7.84e-05	8.54e-01	1	1	1	9.18e-03	
8	x	1307131	4.95e-05	7.52e-01	1	1	1	1.14e-02	
16	x	72315398	4.92e-05	9.92e-01	1	1	1	1.24e-02	
3	H1FOO	130747968	1.85e-05	8.90e-01	1	1	1	1.65e-02	
12	x	39245441	1.21e-07	4.38e-01	1	1	1	1.66e-02	
7	CCDC132	92588411	1.28e-07	4.09e-01	1	1	1	1.77e-02	
1	x	110389963	1.46e-07	2.59e-01	1	1	1	1.99e-02	
9	x	21342862	7.95e-05	9.45e-01	1	1	1	2.18e-02	
2	PRKCE	46170592	1.78e-05	3.40e-01	1	1	1	2.21e-02	
17	x	52636364	3.45e-05	5.20e-01	1	1	1	2.26e-02	
1	x	82547439	5.51e-05	8.89e-01	1	1	1	2.72e-02	
6	x	156238397	4.73e-05	1.43e-01	1	1	1	2.80e-02	
11	x	61956393	2.16e-06	5.37e-01	1	1	1	4.10e-02	
10	x	135319919	3.90e-05	6.76e-01	1	1	1	4.33e-02	
12	x	66026196	2.57e-06	5.08e-01	1	1	1	4.42e-02	
8	x	25535212	2.46e-05	3.44e-01	1	1	1	5.31e-02	
15	IQGAP1	88817746	8.64e-05	2.22e-01	1	1	1	5.76e-02	
6	SIRT5	13707282	3.98e-05	3.66e-01	1	1	1	6.84e-02	
1	x	70907559	3.96e-05	4.71e-01	1	1	1	7.24e-02	
1	CEP350	176696794	7.14e-05	4.50e-01	1	1	1	1.04e-01	
12	x	8955888	7.85e-06	2.07e-01	1	1	1	1.10e-01	
11	x	94090071	5.22e-05	3.08e-01	1	1	1	1.29e-01	
2	x	4641380	9.57e-05	3.68e-01	1	1	1	1.39e-01	
1	x	23749819	8.10e-05	2.08e-01	1	1	1	1.58e-01	

7	x	105466371	4.61e-05	9.90e-02	1	1	1	2.32e-01
5	x	4489013	8.96e-05	3.83e-02	1	1	1	4.40e-01
1	x	215993345	2.67e-05	1.32e-02	1	1	0.59	4.90e-01

Table S3: Replicability analysis for the study of [21]: GWAS of Crohn's disease. The number of SNPs in the primary study was 635547, and 126 SNPs were followed-up.

row #	Chr.	Position	p1	p2	$(l_{00}, c_2) = (0, 0.5)$	$(l_{00}, c_2) = (0.5, 0.5)$	$(l_{00}, c_2) = (0.8, 0.5)$	p_meta	
1	1	67417979	3.19e-34	1.50e-36	4.05e-28	2.03e-28	8.11e-29	2.15e-68	
2	1	67414547	5.05e-36	3.10e-29	3.91e-27	3.91e-27	3.91e-27	3.33e-63	*
3	1	67387537	1.35e-24	5.62e-17	4.72e-15	4.72e-15	4.72e-15	1.82e-39	
4	2	233962410	5.66e-21	7.67e-14	4.83e-12	4.83e-12	4.83e-12	1.18e-32	*
5	5	40428485	2.51e-22	2.79e-08	1.34e-06	1.14e-06	1.14e-06	3.09e-27	*
6	5	40437266	2.26e-22	3.18e-08	1.34e-06	1.14e-06	1.14e-06	3.41e-27	
7	2	233965368	1.28e-21	3.66e-05	5.76e-04	5.76e-04	5.76e-04	4.61e-25	
8	10	64108492	9.51e-12	1.61e-10	1.73e-06	1.14e-06	4.84e-07	2.23e-20	*
9	5	131798704	2.29e-09	3.52e-11	2.08e-04	1.04e-04	4.16e-05	1.16e-18	*
10	18	12769947	5.95e-12	2.41e-07	7.20e-06	6.48e-06	6.48e-06	2.55e-17	*
11	10	101281583	8.53e-11	1.69e-07	1.05e-05	6.48e-06	5.32e-06	1.53e-16	*
12	5	150239060	3.18e-11	2.57e-07	7.20e-06	6.48e-06	6.48e-06	1.70e-16	*
13	10	101282445	9.09e-11	3.10e-07	1.05e-05	7.10e-06	7.10e-06	3.05e-16	
14	18	12799340	3.27e-11	1.23e-06	2.38e-05	2.38e-05	2.38e-05	7.05e-16	
15	5	150203580	4.09e-11	7.47e-07	1.57e-05	1.57e-05	1.57e-05	7.33e-16	
16	13	43355925	8.04e-08	1.33e-07	5.68e-03	2.84e-03	1.20e-03	1.04e-13	*
17	5	158747111	4.40e-09	3.66e-06	3.73e-04	1.86e-04	7.46e-05	1.93e-13	*
18	6	167408399	1.65e-07	3.26e-07	8.74e-03	4.56e-03	1.91e-03	5.22e-13	*
19	3	49696536	1.08e-07	5.64e-07	6.54e-03	3.27e-03	1.37e-03	5.76e-13	*
20	17	37767727	2.97e-06	9.15e-08	8.58e-02	4.49e-02	1.99e-02	3.41e-12	*
21	3	49676987	9.47e-08	2.24e-06	6.02e-03	3.01e-03	1.27e-03	3.55e-12	
22	1	197667523	3.41e-07	2.34e-06	1.44e-02	7.47e-03	3.33e-03	7.17e-12	*
23	12	39104262	8.95e-08	6.55e-05	5.99e-03	2.99e-03	1.26e-03	6.36e-11	
24	6	106541962	1.85e-06	7.70e-06	6.03e-02	3.18e-02	1.38e-02	1.22e-10	*
25	9	114645994	1.96e-07	6.58e-05	9.46e-03	4.91e-03	2.17e-03	1.30e-10	*
26	12	38888207	6.64e-08	1.65e-04	4.96e-03	2.48e-03	1.91e-03	1.54e-10	*
27	6	20836710	1.26e-07	2.78e-04	7.28e-03	3.64e-03	2.92e-03	4.48e-10	*
28	11	75978964	7.16e-08	7.32e-04	8.02e-03	6.83e-03	6.36e-03	6.60e-10	*
29	21	44439989	5.41e-06	1.59e-05	1.32e-01	7.02e-02	3.06e-02	7.04e-10	*
30	1	157665119	1.75e-07	4.81e-04	8.90e-03	4.91e-03	4.33e-03	7.30e-10	*
31	1	169593891	2.01e-07	3.21e-04	9.46e-03	4.91e-03	3.24e-03	7.66e-10	*
32	1	197691964	9.69e-07	1.00e-04	3.52e-02	1.92e-02	7.95e-03	8.10e-10	
33	10	35327656	4.24e-06	2.53e-05	1.10e-01	5.86e-02	2.51e-02	8.93e-10	*
34	19	1074378	5.80e-09	3.47e-03	2.82e-02	2.57e-02	2.19e-02	1.06e-09	
35	19	1075031	6.48e-09	2.10e-02	1.10e-01	9.80e-02	8.97e-02	1.18e-09	
36	20	61798026	7.60e-07	1.38e-04	2.93e-02	1.56e-02	6.44e-03	1.30e-09	
37	7	50081722	1.58e-05	9.41e-06	2.73e-01	1.54e-01	7.07e-02	1.39e-09	
38	6	167405736	1.65e-07	1.21e-03	1.09e-02	1.02e-02	9.24e-03	1.58e-09	
39	9	4971602	3.40e-07	4.30e-04	1.44e-02	7.47e-03	4.01e-03	1.73e-09	*
40	6	32789255	1.53e-08	3.82e-03	2.93e-02	2.75e-02	2.35e-02	2.17e-09	
41	8	126609233	2.45e-06	1.09e-04	7.41e-02	3.80e-02	1.68e-02	2.25e-09	*
42	7	50046933	2.46e-05	1.10e-05	3.68e-01	2.14e-01	1.04e-01	2.30e-09	*
43	17	35294289	1.06e-06	2.92e-04	3.74e-02	2.04e-02	8.42e-03	2.50e-09	*
44	6	32484449	7.23e-09	6.02e-03	4.10e-02	3.79e-02	3.20e-02	2.60e-09	
45	8	126603853	1.90e-06	1.82e-04	6.04e-02	3.18e-02	1.38e-02	2.78e-09	
46	9	114648320	1.31e-07	4.22e-03	3.13e-02	2.95e-02	2.51e-02	3.67e-09	
47	21	15727091	1.03e-05	4.58e-05	2.02e-01	1.11e-01	4.85e-02	3.70e-09	*
48	1	114015850	7.75e-06	8.25e-05	1.67e-01	9.29e-02	3.94e-02	4.95e-09	
49	1	114089610	9.05e-06	1.01e-04	1.89e-01	1.05e-01	4.42e-02	7.30e-09	*
50	10	35589263	6.05e-06	1.76e-04	1.40e-01	7.69e-02	3.20e-02	8.04e-09	
51	21	44436378	5.21e-06	3.61e-04	1.30e-01	7.02e-02	3.01e-02	1.43e-08	
52	21	15734423	1.00e-05	4.44e-04	2.02e-01	1.11e-01	4.80e-02	3.36e-08	
53	3	49499240	2.42e-08	1.94e-01	5.28e-01	5.28e-01	5.28e-01	3.56e-08	
54	9	4978761	1.96e-06	1.62e-03	6.08e-02	3.19e-02	1.38e-02	4.34e-08	
55	2	61129193	3.07e-06	2.80e-03	8.67e-02	4.54e-02	2.00e-02	6.36e-08	
56	1	169594596	1.90e-07	2.60e-02	1.30e-01	1.11e-01	1.07e-01	9.01e-08	
57	3	49425868	2.84e-08	1.07e-01	3.28e-01	3.16e-01	3.16e-01	1.20e-07	
58	13	43497789	6.90e-07	8.82e-03	5.85e-02	5.05e-02	4.36e-02	1.44e-07	
59	2	61098480	3.82e-06	5.65e-03	1.03e-01	5.40e-02	3.10e-02	1.57e-07	
60	6	20797924	1.83e-07	2.88e-02	1.34e-01	1.19e-01	1.15e-01	1.64e-07	
61	6	5096246	3.54e-07	1.92e-02	1.03e-01	9.29e-02	8.34e-02	3.48e-07	
62	1	157691986	2.98e-07	2.77e-02	1.32e-01	1.16e-01	1.13e-01	3.82e-07	

63	17	29611838	2.01e-06	1.35e-02	7.91e-02	7.02e-02	6.19e-02	5.34e-07
64	17	37824128	7.42e-06	7.40e-03	1.65e-01	9.25e-02	3.85e-02	7.10e-07
65	19	18300383	5.43e-08	5.26e-02	2.02e-01	1.92e-01	1.84e-01	7.54e-07
66	2	27652888	3.62e-05	3.81e-03	5.06e-01	2.88e-01	1.37e-01	1.15e-06
67	6	3378317	1.04e-06	3.91e-02	1.67e-01	1.54e-01	1.45e-01	1.37e-06
68	2	102521887	1.02e-05	1.60e-02	2.02e-01	1.11e-01	7.07e-02	1.45e-06
69	2	27642591	3.44e-05	1.08e-02	4.86e-01	2.80e-01	1.33e-01	2.30e-06
70	2	230934834	7.59e-06	5.44e-02	2.02e-01	1.93e-01	1.85e-01	2.48e-06
71	20	61820069	2.04e-07	3.30e-01	7.58e-01	7.58e-01	7.58e-01	2.66e-06
72	6	3379241	1.15e-06	5.82e-02	2.13e-01	2.04e-01	1.96e-01	2.83e-06
73	10	75302766	1.23e-05	3.14e-02	2.23e-01	1.28e-01	1.24e-01	3.03e-06
74	1	7840274	1.47e-06	5.41e-02	2.02e-01	1.93e-01	1.85e-01	3.63e-06
75	6	149618772	3.64e-06	4.40e-02	1.85e-01	1.67e-01	1.59e-01	4.39e-06
76	6	21578398	4.97e-06	6.78e-02	2.41e-01	2.31e-01	2.25e-01	5.02e-06
77	22	20264229	1.25e-06	3.25e-01	7.58e-01	7.58e-01	7.58e-01	6.26e-06
78	11	63906946	4.74e-06	2.45e-01	6.30e-01	6.30e-01	6.30e-01	7.44e-06
79	4	187576360	1.35e-06	8.65e-02	2.87e-01	2.80e-01	2.76e-01	7.81e-06
80	2	230916728	8.93e-06	8.43e-02	2.83e-01	2.80e-01	2.72e-01	9.04e-06
81	17	29849794	1.25e-05	9.61e-02	3.10e-01	2.99e-01	2.99e-01	1.01e-05
82	2	102529086	1.08e-05	4.93e-02	2.02e-01	1.83e-01	1.75e-01	1.11e-05
83	20	57351084	1.73e-06	1.01e-01	3.22e-01	3.10e-01	3.10e-01	1.18e-05
84	4	187585769	1.34e-06	1.07e-01	3.28e-01	3.16e-01	3.16e-01	1.33e-05
85	16	84545499	4.74e-06	2.26e-01	5.87e-01	5.87e-01	5.87e-01	1.40e-05
86	18	17927329	1.59e-05	4.43e-02	2.73e-01	1.67e-01	1.59e-01	1.44e-05
87	18	54054001	5.56e-06	2.07e-01	5.55e-01	5.55e-01	5.55e-01	1.97e-05
88	14	75071147	4.71e-06	1.52e-01	4.35e-01	4.26e-01	4.26e-01	2.25e-05
89	5	37949301	1.74e-06	2.73e-01	6.68e-01	6.68e-01	6.68e-01	2.41e-05
90	10	75324937	1.12e-05	1.04e-01	3.28e-01	3.16e-01	3.16e-01	3.32e-05
91	6	21565929	1.09e-05	1.23e-01	3.68e-01	3.56e-01	3.56e-01	3.40e-05
92	11	63967228	1.60e-05	8.82e-02	2.89e-01	2.81e-01	2.78e-01	3.45e-05
93	12	58059725	2.84e-05	1.49e-01	4.32e-01	4.22e-01	4.22e-01	3.97e-05
94	22	20281207	8.65e-07	4.93e-01	1.00e+00	1.00e+00	9.94e-01	4.55e-05
95	4	106463957	6.25e-06	2.71e-01	6.68e-01	6.68e-01	6.68e-01	5.03e-05
96	1	222692358	2.73e-06	3.93e-01	8.46e-01	8.46e-01	8.39e-01	5.08e-05
97	4	7649390	3.24e-06	3.52e-01	7.99e-01	7.99e-01	7.92e-01	5.27e-05
98	17	35315722	3.41e-06	4.19e-01	8.95e-01	8.95e-01	8.87e-01	5.45e-05
99	3	133674827	6.84e-06	1.61e-01	4.56e-01	4.46e-01	4.46e-01	6.21e-05
100	1	7766478	1.45e-05	2.11e-01	5.60e-01	5.60e-01	5.60e-01	6.82e-05
101	8	83235127	1.32e-05	2.23e-01	5.85e-01	5.85e-01	5.85e-01	1.28e-04
102	21	39215894	8.73e-06	2.63e-01	6.63e-01	6.63e-01	6.63e-01	1.65e-04
103	10	122495603	2.08e-05	3.63e-01	8.10e-01	8.10e-01	8.02e-01	1.98e-04
104	14	75056332	1.28e-05	3.31e-01	7.58e-01	7.58e-01	7.58e-01	2.22e-04
105	13	80961793	1.61e-07	3.72e-01	8.22e-01	8.22e-01	8.15e-01	2.23e-04
106	18	75866208	1.38e-06	2.56e-01	6.52e-01	6.52e-01	6.52e-01	2.80e-04
107	12	13070503	8.89e-06	4.35e-01	9.18e-01	9.18e-01	9.11e-01	3.27e-04
108	10	132842492	2.65e-05	4.41e-01	9.18e-01	9.18e-01	9.11e-01	3.88e-04
109	5	37948752	1.06e-05	4.41e-01	9.18e-01	9.18e-01	9.11e-01	4.78e-04
110	13	80973593	2.64e-07	3.29e-02	1.48e-01	1.32e-01	1.28e-01	5.54e-04
111	12	13046606	4.05e-05	4.55e-01	9.40e-01	9.40e-01	9.32e-01	7.51e-04
112	10	1453158	7.72e-06	2.90e-01	6.96e-01	6.96e-01	6.96e-01	7.56e-04
113	1	181883035	1.04e-05	3.88e-01	8.43e-01	8.43e-01	8.36e-01	9.32e-04
114	18	59311578	1.01e-05	4.98e-01	1.00e+00	1.00e+00	9.96e-01	9.48e-04
115	7	130385443	1.24e-05	3.81e-01	8.35e-01	8.35e-01	8.28e-01	9.64e-04
116	18	55030807	1.75e-05	3.04e-01	7.23e-01	7.23e-01	7.23e-01	1.06e-03
117	19	50999246	1.95e-05	4.60e-01	9.42e-01	9.42e-01	9.35e-01	1.32e-03
118	15	72660732	7.44e-06	2.73e-01	6.68e-01	6.68e-01	6.68e-01	1.33e-03
119	18	55028896	8.34e-06	3.56e-01	8.01e-01	8.01e-01	7.94e-01	1.80e-03
120	16	84542932	3.44e-04	2.78e-01	1.00e+00	1.00e+00	7.88e-01	2.39e-03
121	8	107779719	2.92e-05	3.14e-01	7.40e-01	7.40e-01	7.40e-01	2.78e-03
122	12	58052436	1.14e-05	2.89e-01	6.96e-01	6.96e-01	6.96e-01	3.45e-03
123	18	54054701	9.40e-06	1.95e-01	5.28e-01	5.28e-01	5.28e-01	5.28e-03
124	18	75865061	2.11e-06	1.08e-01	3.28e-01	3.16e-01	3.16e-01	6.92e-03
125	15	72685472	5.81e-06	7.71e-02	2.70e-01	2.59e-01	2.52e-01	1.27e-02
126	8	107743073	2.59e-05	1.43e-01	4.19e-01	4.09e-01	4.09e-01	1.36e-02