Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study

Ruth Heller[a,1,2], Marina Bogomolov[b,1], and Yoav Benjamini[a,c]

[a]Department of Statistics and Operations Research and [c]The Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv 6997801, Israel; and [b]Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 3200003, Israel

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved September 25, 2014 (received for review August 8, 2013)

We propose a formal method to declare that findings from a primary study have been replicated in a follow-up study. Our proposal is appropriate for primary studies that involve large-scale searches for rare true positives (i.e., needles in a haystack). Our proposal assigns an *r* value to each finding; this is the lowest false discovery rate at which the finding can be called replicated. Examples are given and software is available.

false discovery rate | genome-wide association studies | metaanalysis | multiple comparisons | *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) examined for associations with disease, genes examined for differential expression, pathways examined for enrichment, and protein pairs examined for 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 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. Replicability has been the cornerstone of science as we know it since the foundation of experimental science. Possibly the first documented example is the discovery of a phenomenon related to vacuum, made by Huygens in Amsterdam in the 17th century, who traveled to Boyle's laboratory in London to replicate the experiment and prove that the scientific phenomenon was not idiosynchronic to his specific laboratory with his specific equipment (1). In modern research, the lack of replicability has deeply bothered behavioral scientists that compare the behavior of different strains of mice, e.g., in knockout experiments. It is well documented that in different laboratories, the comparison of behaviors of the same two strains may lead to opposite conclusions that are both statistically significant (refs. 2, 3, and 4, chap. 4). An explanation may be the different laboratory environment (i.e., personnel, equipment, measurement techniques) affecting differently the study strains (i.e., an interaction of strain with laboratory). This means that the null hypothesis that the effect is, say, nonpositive is true in one laboratory, but false in the other laboratory, and thus the positive effect is not replicated in both laboratories. In genomic research, the interest is in the genetic effect on phenotype. In different studies of the same associations with phenotype, we seem to be testing the same hypotheses but the hypotheses tested are actually much more particular. Whether a hypothesis is true may depend on the cohorts in the study that are from specific populations exposed to specific environments (for particular examples, see *Results*). However, if discoveries are made, it is of great interest to see whether these discoveries are replicated in different cohorts, from different populations, with different environmental exposures and different measurement techniques. The paramount importance of having replicated findings is well recognized in genomic research (5). In particular, this is so in genome-wide association studies (GWAS) (6, 7). As noted in ref. 8, 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, it is important to observe the same association in other studies using similar, but not identical, subpopulations and methods. Obviously, splitting the data from the same study into two independent parts and doing the same analysis on each does not answer the above concerns.

Replicability problems arise in many additional scientific areas, and discussions of these problems reached prominent general-interest venues, for instance refs. 9 and 10. We need to have an objective way to declare that a certain study really replicates the findings in another study. This paper makes a concrete, objective, easy to apply, and rigorously motivated way to determine that a finding has been replicated.

Replicability vs. Metaanalysis

In many areas it is common to combine the results of studies that examine the same scientific hypotheses by a metaanalysis. Pooling results across studies is especially attractive when single studies are underpowered, using the potential increase in power of combining the studies, but the metaanalysis P value tests only the null hypothesis of no signal in all studies. As a result, a strong signal in one of the studies (with P value close to zero) is enough to declare the metaanalysis finding highly significant. In statistical terms, denoting the hypothesis that there is no real effect for feature j in study i by $H_{0i}(j)$, the metaanalysis P value tests the intersection null hypothesis $\cap_i H_{0i}(j)$ that there is no effect in all studies. Rejection of the intersection hypothesis establishes that in at least one study

Significance

The use of big data is becoming a central way of discovering knowledge in modern science. Large amounts of potential findings are screened to discover the few real ones. To verify these discoveries a follow-up study is often conducted, wherein only the promising discoveries are followed up. Such follow-up studies are common in genomics, in proteomics, and in other areas where high-throughput methods are used. We show how to decide whether promising findings from the preliminary study are replicated by the follow-up study, keeping in mind that the preliminary study involved an extensive search for rare true signal in a vast amount of noise. The proposal computes a number, the *r* value, to quantify the strength of replication.

Author contributions: R.H., M.B., and Y.B. designed research; R.H. and M.B. performed research; R.H. and M.B. contributed new reagents/analytic tools; R.H. and M.B. analyzed data; and R.H., M.B., and Y.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]R.H. and M.B. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: ruheller@post.tau.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1314814111/-/DCSupplemental.

Table 1. Replicability analysis for FDR control for the study of ref. 11: GWAS of IgA nephropathy in Han Chinese

Chr.	Position	Gene	p1	p2	p_meta	*r* value
6	32,685,358	HLA-DRB1	8.19e-08	8.57e-14	4.13e-20	0.0074
8	6,810,195	DEFAs	2.04e-07	1.25e-07	3.18e-14	0.0090
6	32,779,226	HLA-DQA/B	3.28e-08	3.57e-06	3.43e-13	0.0059
22	28,753,460	MTMR3	2.30e-07	2.02e-05	1.17e-11	0.0090
6	30,049,922	HLA-A	4.05e-09	3.68e-04	1.74e-11	0.0090
17	7,403,693	TNFSF13	1.50e-06	2.52e-05	9.40e-11	0.0413
17	7,431,901	MPDU1	5.52e-07	3.16e-04	4.31e-10	0.0169

The number of SNPs in the primary study was 444,882, and 61 were followed up. For the seven most significant metaanalysis *P* values, the position (columns 1–3), the primary and follow-up study *P* values (columns 4 and 5), the metaanalysis *P* values (column 6), and the *r* values (column 7) are shown. Table S1 shows the results for all 61 SNPs followed up. The lower bound for f_{00} was $l_{00} = 0.8$ for the *r*-value computation.

$H_{0i}(j)$ is false, but possibly only in a single one. Thus, a metaanalysis discovery based on a few studies is no better than a discovery from a single large study in assessing replicability.

In GWAS, a typical table of results reports the *P* values in the primary and follow-up study, side by side, as well as the metaanalysis *P* values, for the SNPs with the smallest metaanalysis *P* values. Table 1 (columns 1–6) is an example of such a table of results (11). In replicability analysis, the null hypothesis of signal in at most one study is tested, the rejection of which yields the statistical significance of the replicability claim. In statistical terms, the replicability claim is established for feature *j* by rejecting the union null hypothesis $H_{01}(j) \cup H_{02}(j)$.

Note that replicability is sometimes referred to as reproducibility, but we prefer to view reproducibility as a property of each single study, a distinction made in ref. 12.

The *r* Value for Replicability

If each study examines only a single hypothesis, and the hypothesis in one study is rejected at the 0.05 level, and the hypothesis in the second study is also rejected in the same direction at the 0.05 level, then replicability is intuitively established. This is also a sound claim, in the sense that the probability of claiming that a finding is replicated if the null hypothesis is true in at least one of the studies is at most 0.05. The need for a statistical framework for establishing replicability becomes essential with the use of high-throughput methods. The potential to err in inference when more than one study is involved is more severe when each study is examining simultaneously many features. The choices for selection are much wider. Therefore, the statistical methods needed are more complicated than the very intuitive statistical method for establishing replicability when a single feature is involved.

Multiple-testing methods are widely used to adjust for the effect of selection, either by controlling the probability of erroneously selecting even a single feature [i.e., the family-wise error rate (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. Our method reports the *r* value that can be defined for either error rate for replicability analysis. Here we emphasize the FDR:

Definition: The FDR *r* value for feature *i* is the lowest FDR level at which we can say that the finding is among the replicated ones.

The smaller the *r* value is, the stronger the evidence in favor of replicability of the finding. It can be compared with any desired level of FDR in the same way that a *P* value is commonly compared with the desired false detection parameter α.

In this work we introduce a method for computing *r* values for features examined in primary and follow-up studies. We suggest to complement tables of results that report for selected findings the primary, follow-up, and metaanalysis *P* values with an additional column of *r* values. The *r* values in column 7 of Table 1 are all below 0.05, concurring with the main replicability findings of ref. 11. The ranking of *r* values is different from the ranking of the metaanalysis *P* values, indicating the novelty of the added information. Table 2 shows the results of a somewhat more complicated example discussed below, where the difference between the metaanalysis and the replicability conclusions is more dramatic.

Assessing Replicability from Follow-up Studies

We concentrate on the widely used design in "omics" that examines *m* features in the primary study and only a fraction thereof in the follow-up study. For other designs, see *Assessing Replicability in Other Designs*.

When $m = 1$, as we discussed above, replicability is established at the 0.05 significance level if both *P* values are at most 0.05. When $m > 1$, the multiplicity of features should be taken into account. Note that replicability cannot be assessed by the following common practice: Features are screened in a primary study, then the features with promising results are followed, and then the discoveries are based only on a testing procedure appropriate for the single follow-up study. This is so for two reasons: first, because screening is typically done without appropriate control over false positives (examples in *Results*); and second, even if the screening procedure controls the false positives at a level appropriate for the single primary study, this level needs to be further adjusted. Otherwise, applying a testing procedure appropriate for the single follow-up study is not enough to offer control over false replicability claims.

A simple approach can be to apply a multiple-testing procedure on the maximum of the two studies' *P* values, setting conservatively the maximum value at one if the feature was not followed up. This is not recommended because the price paid for multiplicity is too large. More powerful procedures for FWER and FDR control were suggested for this design in ref. 13, in which effectively the primary study *P* values have to be adjusted for the multiplicity of *m* hypotheses, but the follow-up study *P* values need to be adjusted only for the multiplicity of the hypotheses followed up. Here we suggest a generalization of the method of ref. 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. We demonstrate our proposal on *P* values from GWAS.

Table 2. Replicability analysis for FDR control for the study of ref. 17 on GWAS of T2D

Chr.	Position	p.primary	p1	p2	p_meta	*r* value
7	27,953,796	1.55e-04	8.07e-05	1.34e-07	4.96e-14	0.0055
10	12,368,016	4.21e-04	5.40e-05	1.49e-04	1.21e-10	0.0055
12	69,949,369	1.80e-05	9.83e-03	4.35e-05	1.11e-09	0.1490
2	43,644,474	1.83e-04	1.62e-03	9.22e-05	1.12e-09	0.0441
3	64,686,944	5.44e-04	1.02e-04	3.47e-03	1.17e-08	0.0254
1	120,230,001	1.14e-04	2.89e-03	1.95e-03	4.10e-08	0.0604
12	53,385,263	3.18e-05	3.11e-03	8.81e-03	1.79e-07	0.0604
3	12,252,845	1.05e-05	4.50e-03	1.22e-02	1.97e-07	0.0765
1	120,149,926	1.35e-03	1.17e-03	7.84e-03	4.04e-07	0.0431
6	43,919,740	5.41e-05	1.46e-03	9.49e-02	4.03e-06	0.2090
2	60,581,582	3.38e-05	1.38e-03	6.54e-01	1.02e-04	1.0000

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 and 2), the primary study *P* values and first and second follow-up studies *P* values (columns 3–5), the metaanalysis *P* values from all three studies (column 6), and the *r* values quantifying the evidence of replicability from the first to the second follow-up study (column 7) are shown. The lower bound for f_{00} was $l_{00} = 0$ for the *r*-value computation, because the set of SNPs in the first follow-up study is already believed to be associated with T2D.

However, the P values can obviously come from other applications such as exome-sequencing studies, ChIP experiments, or microarray studies.

Let f_{00} denote the fraction of features, of the m features examined in the primary study, that are null in both studies. We cannot estimate f_{00} from the data, because only a handful of promising features (SNPs) are followed up in practice. However, f_{00} is typically closer to one than to zero, and we can give a conservative guess for a lower bound on f_{00}, call it l_{00}. In typical GWAS on the whole genome, $l_{00}=0.8$ is conservative. We can exploit the fact that $l_{00}>0$ to gain power.

Computation of *r* Values for FDR Replicability.

i) Data input:

a) m, the number of features examined in the primary study.

b) $\mathcal{R}_1$, the set of features selected for follow-up based on primary study results. Let $R_1=|\mathcal{R}_1|$ be their number.

c) $\{(p_{1j},p_{2j}): j\in\mathcal{R}_1\}$, where p_{1j} and p_{2j} are, respectively, the primary and follow-up study P values for feature $j\in\mathcal{R}_1$.

ii) Parameters input:

a) $l_{00}\in[0,1)$, the lower bound on f_{00} (see above); default value for whole-genome GWAS is $l_{00}=0.8$.

b) $c_2\in(0,1)$, the emphasis given to the follow-up study (*Variations* section); default value is $c_2=0.5$.

iii) Definition of the functions $f_i(x), i\in\mathcal{R}_1, x\in(0,1)$:

a) Compute $c_1=(1-c_2)/(1-l_{00}(1-c_2x))$.

b) For every feature $j\in\mathcal{R}_1$ compute the following e values:

$$e_j=\max\left(\frac{1}{c_1}p_{1j},\frac{R_1}{mc_2}p_{2j}\right), j\in\mathcal{R}_1.$$

c) Let $f_i(x)=\min_{\{j:e_j\geq e_i, j\in\mathcal{R}_1\}}(e_jm/\text{rank}(e_j))$, where $\text{rank}(e_j)$ is the rank of the e value for feature $j\in\mathcal{R}_1$ (with maximum rank for ties).

iv) The FDR r value for feature $i\in\mathcal{R}_1$ is the solution to $f_i(r_i)=r_i$ if a solution exists in $(0,1)$ and 1 otherwise. The solution is unique; see *SI Text, Lemma S1.1* for a proof.

The r values can be computed using our web application, which is available in RStudio (spark.rstudio.com/shayy/radjust). An R script is also available in RunMyCode, www.runmycode.org/companion/view/542.

The adjustment in step *iiic* is similar to the computation of the adjusted P values (14) for the Benjamini–Hochberg (BH) procedure (15), the important difference being that e values are used instead of P values. The replicability claims at a prefixed level q, say $q=0.05$, are all indexes with r values at most 0.05. The FDR for replicability analysis is then controlled at level 0.05; details are in *Derivation and Properties*.

For $l_{00}=0$, declaring as replicated the findings with r values at most q coincides with procedure 3.2 in ref. 13. It is easy to see that with $l_{00}>0$, we will have at least as many replicability claims as with procedure 3.2 in ref. 13. Next we show in GWAS examples and simulations that the power increases with l_{00} and can lead to many more discoveries than with procedure 3.2 in ref. 13, while maintaining FDR control.

Results. 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. In these examples, the primary and follow-up studies differ in the subpopulations examined and may also differ in design and analysis. In addition, the primary and follow-up studies may differ in quality. It is therefore of scientific importance to discover which associations were replicated. The examples differ in design and in the selection rules for forwarding SNPs for follow-up. In the first example, there is one primary study and one follow-up study, a 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 metaanalysis of three studies, more than a hundred hypotheses are followed-up, and a few dozen SNPs have r values below 0.05. In the third example, there are three stages: a primary study, then a follow-up study, and then an additional follow-up study that is based on the first follow-up study.

Our first example is GWAS of IgA nephropathy in Han Chinese (11). To discover association between SNPs and IgA nephropathy, 444,882 SNPs were genotyped in 1,523 cases from southern China and 4,276 controls from Singapore and from southern and northern China, with the same ancestral origin. For follow-up, 61 SNPs were measured in two studies: 1,402 cases and 1,716 controls from northern China and 1,301 cases and 1,748 controls from southern China. The 61 SNPs selected for follow-up had primary study P values below 10^{-5}. Table 1 shows the 7 SNPs with the smallest metaanalysis P values, of the 61 SNPs followed up. The associations for these 7 SNPs have been replicated with r values ≤ 0.05 for $l_{00}=0.8$. The 7 SNPs clearly stand out from the remaining 54 SNPs followed up that have r values of 1 (Table S1). If the researcher is willing to assume only a lower bound of 0.5 or of 0 for f_{00}, then the r values are larger than with $l_{00}=0.8$. Table S1 shows that with $l_{00}=0.5$ and $l_{00}=0$, respectively, only 6 and 5 SNPs had r values below 0.05.

Our second example is GWAS of Crohn's disease (CD). To discover associations between SNPs and CD (16), 635,547 SNPs were examined in 3,230 cases and 4,829 controls of European descent, collected in three separate studies: NIDDK4, WTCCC5, and a Belgian–French study. For follow-up, 126 SNPs were measured in 2,325 additional cases and 1,809 controls as well as in an independent family-based dataset of 1,339 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 S2 shows the 126 SNPs followed up. Applying our proposal with parameter $l_{00}=0.8$, we decide that 52 SNPs have replicated associations at r values ≤ 0.05. The 52 SNPs with replicated associations did not correspond to the 52 SNPs with the smallest metaanalysis P values. For example, the SNP in row 35 had the 35th smallest metaanalysis P value, but its r value was 0.09, and thus it was not among the 52 replicated discoveries. The last column of Table S2 marks 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 1 of the main text of ref. 16. These 30 SNPs have r values below 0.05, so they are a subset of the 52 replicated discoveries. Our replicability analysis discovers more loci, in particular three loci (rows 34, 44, and 59 in Table S2) that did not reach the conservative Bonferroni threshold of ref. 16 in the follow-up study P values, yet were pointed out in table 2 of ref. 16 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 (17), more than 2 million SNPs were imputed from about 400,000 SNPs collected in 4,549 cases and 5,579 controls combined from three separate studies: DGI, WTCCC, and FUSION. For follow-up, 68 SNPs were measured in 10,037 cases and 12,389 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 of the 68 SNPs were measured in 14,157 cases and 43,209 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 metaanalysis P values below 10^{-5} when combining the evidence from the primary study and the first follow-up study. Although 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 2 shows the 11 SNPs followed up from the first follow-up study to the second follow-up study. Applying our proposal with $l_{00}=0$, we decide that 5 SNPs have replicated associations with r-values ≤ 0.05. Note that we set $l_{00}=0$ because most of the 68 SNPs in the first follow-up study are already believed to be associated with the disease.

Derivation and Properties

Here we give the formal framework for replicability analysis and the theoretical properties of our proposal. The family of m features examined in the primary study, indexed by $I=\{1,\ldots,m\}$, may be divided into four subfamilies with the following indexes: I_{00}, I_{01}, I_{10}, and I_{11}, for the features with hypotheses that are, respectively, null in both studies, null in the primary study only, null in the follow-up study only, and nonnull in both studies. Suppose R replicability claims are made by an analysis. Denoting by R_{ij} the number of replicability claims from subfamily 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.

The FDR for replicability analysis is the expected proportion of false replicability claims among all those called replicated:

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

Definition: A stable selection rule satisfies the following condition: for any $j\in\mathcal{R}_1$, fixing all primary study P values except for p_{1j} and changing p_{1j} so that j is still selected, will not change the set $\mathcal{R}_1$.

Stable selection rules include selecting the hypotheses with P values below a certain cutoff or by a nonadaptive multiple-testing procedure on the primary study P values, such as the BH procedure for FDR control or the Bonferroni procedure for FWER control, or selecting the k hypotheses with the smallest P values, where k is fixed in advance.

Theorem 1. *A procedure that declares findings with r values at most q as replicated controls the FDR for replicability analysis at a level at most q if the rule by which the set* $\mathcal{R}_1$ *is selected is a stable selection rule,* $l_{00}\leq f_{00}$, *and the P values within the follow-up study are jointly independent or are positive regression dependent on the subset of true null hypotheses* (*property PRDS*) *and are independent of the primary study P values, in one of the following situations*:

i) *The P values within the primary study are independent.*
ii) *There is arbitrary dependence among the P values within the primary study, when in step iii m is replaced by* $m^*=m\sum_{i=1}^{m}1/i$.
iii) *There is arbitrary dependence among the P values within the primary study, and the selection rule is such that the primary study P values of the features that are selected for follow-up are at most a fixed threshold* $t\in(0,1)$, *when* c_1 *computed in step iiia is replaced by*

$$\tilde{c}_1(x)=\max\left\{a:a\left(1+\sum_{i=1}^{\lceil tm/(ax)-1\rceil}\frac{1}{i}\right)=c_1(x)\right\},$$

where $c_1(x)=(1-c_2)/(1-l_{00}(1-c_2x))$. *Steps iiib and iiic remain unchanged. In step iv, the FDR r value for feature* $i\in\mathcal{R}_1$ *is* $r_i=\min\{x:f_i(x)\leq x\}$ *if a solution exists in* $(0,1)$ *and* 1 *otherwise.*

See *SI Text*, section S1, for a proof. The implication of item *iii* is that for FDR replicability at level q, if $t\leq c_1(q)q/m$, no modification is required, so the procedure that declares as replicated all features with r values at most q controls the FDR at level q on replicability claims for any type of dependency in the primary study. Note that the modification in item *iii* will lead to more discoveries than the modification in item *ii* only if $t<c_1(q)q/(1+\sum_{i=1}^{m-1}1/i)$.

In *SI Text*, section S5, we show realistic GWAS simulations that preserve the dependency across P values in each study. For $l_{00}\in\ \{0,0.8,0.9,0.95,0.99\}$, the FDR of the procedure that declares findings with r values (computed in steps *i*–*iv* of the original proposal) at most 0.05 as replicated is controlled below level 0.05, suggesting that this procedure is valid for the type of dependency that occurs in GWAS. Because this procedure can be viewed as a 2D variant of the BH procedure, and the BH procedure is known to be robust to many types of dependencies, we conjecture that for $l_{00}\leq f_{00}$, our procedure 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 $c_1(q)q/m$ are followed up. In Table S3 we further show the superior power of our procedure over applying the BH procedure on the maximum of the two studies' P values (at level $0.05/(1-l_{00})$, where the maximum value is set to 1 for $j\notin\mathcal{R}_1$).

Variations

Choice of Emphasis Between the Studies. The e-value computation requires combining the P values from the primary and the follow-up study, using a parameter c_2, which we set to be $c_2=0.5$ in the computation above. More generally, for FDR control we need to first select $c_2\in(0,1)$. We shall show the effect the choice of c_2 has on the r values for given P values and argue from power considerations that the choice $c_2=0.5$ is reasonable.

The following procedure is identical to that of declaring the set of findings with r values at most q as replicated; see proof in *SI Text, Lemma S1.1*. First, compute the number of replicability claims at level q as follows:

$$R_2\triangleq\max\left\{r:\sum_{j\in\mathcal{R}_1}\mathbf{I}\left[(p_{1j},p_{2j})\leq\left(\frac{r}{m}c_1(q)q,\frac{r}{R_1}c_2q\right)\right]=r\right\}.$$

Next, declare as replicated findings the set

$$\mathcal{R}_2=\left\{j:(p_{1j},p_{2j})\leq\left(\frac{R_2}{m}c_1(q)q,\frac{R_2}{R_1}c_2q\right),j\in\mathcal{R}_1\right\}.$$

From this equivalent procedure it is clear that a larger choice $c_2\in(0,1)$ will make the threshold that p_{2j} has to pass larger, but the threshold that p_{1j} has to pass smaller, so for the extreme choice $c_2\approx 1$, the discovered findings can only be features with tiny primary study P values, and for the extreme choice of $c_2\approx 0$, the discovered findings can only be features with tiny follow-up study P values. For q small, the primary and follow-up study P values will have the same threshold if $(1/m)((1-c_2)/(1-l_{00}))=c_2/R_1$; i.e., $c_2=1/(1+m(1-l_{00})/R_1)$, which is close to zero if R_1/m is very small (as is typical in GWAS). Therefore, this choice is not recommended unless the power of the follow-up study is extremely large. For the choice $c_2=0.5$, the threshold for the follow-up study P value is larger than for the primary study P value by approximately the factor $m(1-l_{00})/R_1$, i.e., the ratio of the number of hypotheses that should be adjusted for in the primary study to that in the follow-up study. We show next that this choice is good from efficiency considerations.

In simulations, detailed in *SI Text*, section S4, we observed that for a given l_{00} the optimal c_2, i.e., the choice of c_2 that maximizes power, has only a small gain in power over the choice $c_2=0.5$. We considered $m=1,000$ SNPs, of which $f_{00}=0.9$ had no signal,

$f_{01}=0.025$ had signal only in the follow-up study, $f_{10}=0.025$ had 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 BH procedure at level $c_1(q)q$ on the primary study P values, with $q=0.05$. *SI Text*, section S3 has a discussion of the advantage of this selection rule over other selection rules.

The power increased with l_{00} as well as with π_2. Table S4 shows that the gain in power of using $l_{00}>0$ over $l_{00}=0$ can be large. Fig. S1 shows the average power and the power for at least one true replicability discovery as a function of c_2. Fig. S2 shows the FDR as a function of c_2.

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 is greater (0.2, 0.5, or 0.8 in the above simulations). Therefore, for GWAS on the whole genome, we recommend setting $c_2=0.5$.

FWER Replicability. The FWER criterion,

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

is more stringent than the FDR, yet it may sometimes be desired. We define the FWER r value as the lowest FWER level at which we can say that the finding has been significantly replicated. The r value can be compared with any desired level of FWER. An FWER-controlling procedure for replicability analysis was suggested in ref. 13: That study 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 $c_1+c_2=1$. If a nonzero lower bound on f_{00} is available, then this lower bound can be used to choose parameters (c_1,c_2) with a sum greater than 1. Specifically, for FWER control using Bonferroni, the data input and parameters input are the same as in our proposal for FDR replicability in steps *i* and *ii*, but the computation in step *iii* is different. For feature $j\in\mathcal{R}_1$,

$$f_j^{\text{Bonf}}(x)=\max\left(\frac{mp_{1j}}{c_1},\frac{|\mathcal{R}_1|p_{2j}}{c_2}\right),\qquad c_1=\frac{1-c_2}{1-l_{00}(1-c_2x)}.$$

The Bonferroni r value for feature j is the solution to $f_j^{\text{Bonf}}(r_j)=r_j$ if a solution exists in $[0,1)$ and 1 otherwise. The replicability claims at a prefixed level α, say $\alpha=0.05$, are all indexes with r values at most 0.05. The FWER for replicability analysis is then controlled at level 0.05; see *SI Text*, section S6, for the proof.

We computed the Bonferroni r values in a GWAS of thyrotoxic periodic paralysis (TPP) (18). In 70 cases and 800 controls from the Hong Kong (southern) Chinese population, 486,782 SNPs were genotyped. Table S5 shows the four most significant SNPs followed up in an additional 54 southern Chinese TPP cases and 400 healthy Taiwanese controls. The associations were successfully replicated with Bonferroni r values far below 0.05, concurring with the claim that “associations for all four SNPs were successfully replicated” (ref. 18, p. 1027).

Assessing Replicability in Other Designs

The concept of the r value is also relevant to the communication of the results of replicability in other designs. If $n>2$ studies examine a single feature, then replicability of the finding in all n studies is established at the 0.05 significance level if the maximum P value is at most 0.05. However, if a weaker notion of replicability is of interest, e.g., that the finding has been replicated in at least two studies, then the evidence toward replicability can be computed as follows. First, for every subset of $n-1$ studies, a metaanalysis P value is computed. Then, replicability in at least two studies is established at the 0.05 significance level if the maximum of the n metaanalysis P values is at most 0.05. This can be generalized to discover whether the finding has been replicated in at least u studies, where $u\in\{2,\ldots,n\}$, as detailed in ref. 19.

If $n\geq 2$ studies examine each of $m>1$ features, then for each $i\in\{1,\ldots,m\}$ the P value for testing for replicability can be computed as above, but instead of comparing each to 0.05, the BH procedure is applied and the discoveries are considered as replicated findings. The procedure was suggested in ref. 20, and for $n=2$ it amounts to using the maximum of the two studies’ P values for each feature in the BH procedure. The power of this procedure may be low when a large fraction of the null hypotheses are true, because the null hypothesis for replicability analysis is not simple, and the BH procedure is applied on a set of P values that may have a null distribution that is stochastically much larger than uniform. The loss of power of multiple-testing procedures can indeed be severe when using overconservative P values from composite null hypotheses (21). An empirical Bayes approach for discovering whether results have been replicated across studies was suggested in ref. 22 and compared with the analysis of ref. 20, concluding that the empirical Bayes analysis discovers many more replicated findings. The accuracy of the empirical Bayes analysis relies on the ability to estimate well the unknown parameters, and thus it is suitable in problems such as GWAS, where each study contains hundreds of thousands of SNPs, and the dependency across SNPs is local, but may not be suitable for applications with a smaller number of features and nonlocal dependency. A method based on relative ranking of the P values to control their “irreproducible discovery rate” was suggested in ref. 23. A list-intersection test to compare top-ranked gene lists from multiple studies to discover the common significant set of genes was suggested in ref. 24.

To summarize, although for $m=1$ there is a straightforward solution for the problem of establishing replicability, once we move away from this simple setting the problem is more complicated. For designs with more than one potential finding, it is very useful to quantify and report the evidence toward replicability by an r value. The r value is a general concept, but the r-value computation depends on the multiple-testing procedure used, which in turn depends on the design of the replicability problem.

Discussion

The r value was coined in the FDR context, in accordance with the commonly used q value (25). We proposed the r value as an FDR-based measure of significance for replicability analysis. We showed in GWAS examples that the smallest metaanalysis P values may not have the strongest evidence toward replicability of association, and we suggest to report the r values in addition to the metaanalysis P values in the table of results.

In practice, the primary study P values are rarely independent. We prove that our main proposal controls the FDR on replicability claims if the primary study P values are independent and suggest modifications of the proposal that are more conservative but have the theoretical guarantee of FDR control for any type of dependency among the primary study P values. From empirical investigations, we conjecture that the conservative modifications in items *ii* and *iii* of *Theorem 1* are unnecessary for the types of dependencies encountered in GWAS. For our second example, of GWAS in CD, applying the more conservative proposal in item *ii* of *Theorem 1* resulted in 34 discoveries.

We saw examples where the primary study was composed of more than one study, and more than one follow-up study was performed. In the present work, we used all of the information from the primary studies for selection for follow-up, and to establish replicability the metaanalysis P values of the primary studies and the metaanalysis P values of the follow-up studies were used. Alternatively, each study can be considered on its own toward establishing replicability, and inference can be based on $r_{u/n}$ values that quantify the evidence that the finding has been replicated in at least u out of n studies for $2 \leq u \leq n$. 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 $2/n$ toward replicability. This problem has been addressed in ref. 19, but as was shown in ref. 13, alternatives along the lines of the procedures suggested here may benefit from increased power.

A referee pointed out that follow-up studies may be designed to give more trustworthy data, using more expensive equipment, e.g., using PCR or fine linkage analysis. If the aim is to detect associations in the follow-up study, then there is no need to combine the evidence from the primary study with that of the follow-up study. However, if the aim is to detect replicated associations, then it may be of interest to have unequal penalties for the error of discovering a finding that is true only in the primary study and the error of discovering a finding that is true only in the follow-up study. Developing procedures that give unequal penalties to these two errors is a challenging and interesting problem for future research, which may be approached by using weights (26).

ACKNOWLEDGMENTS. We thank Shay Yaacoby for help in developing the web application and Neri Kafkafi for useful discussions on the history of replicability in science. We thank the editor and the reviewers for the extensive comments that led to a substantial improvement of the manuscript. The research of Y.B. leading to these results received funding from the European Research Council (ERC) under the European Community's Seventh Framework Programme (FP7/2007-2013) and ERC Grant Agreement n° (PSARPS-294519). The research of R.H. was supported by Israel Science Foundation Grant 2012896.

1. Shapin S, Schaffer S (1985) *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life* (Princeton Univ Press, Princeton).
2. Crabbe JC, Wahlsten D, Dudek BC (1999) Genetics of mouse behavior: Interactions with laboratory environment. *Science* 284(5420):1670–1672.
3. Kafkafi N, Benjamini Y, Sakov A, Elmer GI, Golani I (2005) Genotype-environment interactions in mouse behavior: A way out of the problem. *Proc Natl Acad Sci USA* 102(12):4619–4624.
4. Crusio WE, Sluyter F, Gerlai RT, Pietropaolo S (2013) *Behavioral Genetics of the Mouse: Genetics of Behavioral Phenotypes*, Cambridge Handbooks in Behavioral Genetics, Vol 1.
5. Lander E, Kruglyak L (1995) Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nat Genet* 11(3):241–247.
6. McCarthy MI, et al. (2008) Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat Rev Genet* 9(5):356–369.
7. Chanock SJ, et al.; NCI-NHGRI Working Group on Replication in Association Studies (2007) Replicating genotype-phenotype associations. *Nature* 447(7145):655–660.
8. Kraft P, Zeggini E, Ioannidis JP (2009) Replication in genome-wide association studies. *Stat Sci* 24(4):561–573.
9. Lehrer J (December 13, 2010) The truth wears off. *The New Yorker*.
10. Anonymous (October 19, 2013) Unreliable research: Trouble at the lab. *The Economist*.
11. Yu XQ, et al. (2012) A genome-wide association study in Han Chinese identifies multiple susceptibility loci for IgA nephropathy. *Nat Genet* 44(2):178–182.
12. Peng RD (2009) Reproducible research and Biostatistics. *Biostatistics* 10(3):405–408.
13. Bogomolov M, Heller R (2013) Discovering findings that replicate from a primary study of high dimension to a follow-up study. *J Am Stat Assoc* 108(504):1480–1492.
14. Reiner A, Yekutieli D, Benjamini Y (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19(3):368–375.
15. Benjamini Y, 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.
16. Barrett JC, et al.; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 40(8):955–962.
17. Zeggini E, et al.; Wellcome Trust Case Control Consortium (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 40(5):638–645.
18. Cheung CL, et al. (2012) Genome-wide association study identifies a susceptibility locus for thyrotoxic periodic paralysis at 17q24.3. *Nat Genet* 44(9):1026–1029.
19. Benjamini Y, Heller R (2008) Screening for partial conjunction hypotheses. *Biometrics* 64(4):1215–1222.
20. Benjamini Y, Heller R, Yekutieli D (2009) Selective inference in complex research. *Philos Trans A Math Phys Eng Sci* 367(1906):4255–4271.
21. Dickhaus T (2013) Randomized *p*-values for multiple testing of composite null hypotheses. *J Stat Plan Inference* 143:1968–1979.
22. Heller R, Yekutieli D (2013) Replicability analysis for genome-wide association studies. *Ann Appl Stat* 8(1):481–498.
23. Li Q, Brown J, Huang H, Bickel P (2011) Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* 5(3):1752–1779.
24. Natarajan L, Pu M, Messer K (2012) Exact statistical tests for the intersection of independent lists of genes. *Ann Appl Stat* 6(2):521–541.
25. Storey J (2002) A direct approach to false discovery rates. *J R Stat Soc Ser B Stat Methodol* 64(3):479–498.
26. Benjamini Y, Hochberg Y (1997) Multiple hypotheses testing with weights. *Scand J Stat* 24(3):407–419.