What is the probability of replicating a statistically significant association in genome-wide association studies?

Wei Jiang[1], Jing-Hao Xue[2], and Weichuan Yu[1*]

[1]*Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China*

[2]*Department of Statistical Science, University College London, London WC1E 6BT, U.K.*

Running Title: Estimating Replicability in GWASs

Keywords: GWAS; replication study; replicability

[*] Correspondence and requests for materials should be addressed to W.Y. (Tel.: +852 2358 7054; E-mail: eeyu@ust.hk).

Biographical Notes

Wei Jiang

Wei Jiang is a PhD candidate in Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China. His doctoral research focuses on statistical issues in multi-stage and multiple genome-wide association studies.

Jing-Hao Xue

Jing-Hao Xue is a senior lecturer in the Department of Statistical Science, University College London, U.K.. His research interests involve: statistical classification, high-dimensional data analysis, computer vision and pattern recognition.

Weichuan Yu

Weichuan Yu is an associate professor in the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China. He is interested in computational analysis problems with biological and medical applications.

Abstract

The goal of genome-wide association studies (GWASs) is to discover genetic variants associated with diseases/traits. Replication is a common validation method in GWASs. We regard an association as true finding when it shows significance in both primary and replication studies. A question worth pondering is what is the probability of a primary association (i.e., a statistically significant association in the primary study) being validated in the replication study? This paper systematically reviews the answers to this question from different points of view. Since Bayesian methods can help us integrate out the uncertainty about the underlying effect of the primary association, we will mainly focus on the Bayesian view in this paper. We refer the Bayesian replication probability as the replication rate (RR). We further describe an estimation method for RR which makes use of the summary statistics from the primary study. We can use the estimated RR to determine the sample size of the replication study and to check the consistency between the results of the primary study and those of the replication study. We describe an R-package to estimate and apply RR in GWASs. Simulation and real data experiments show that the estimated RR has good prediction and calibration performance. We also use these data to demonstrate the usefulness of RR. The R-package is available at http://bioinformatics.ust.hk/RRate.html.

1 Introduction

The goal of genome-wide association studies (GWASs) is to detect genetic variants associated with diseases/traits by genotyping single nucleotide polymorphisms (SNPs) in different individuals [1]. Compared to traditional candidate gene studies based on gene function and pathway information [2], GWASs avoid selection bias by genotyping a dense set of SNPs across the whole genome. Also, GWASs are more powerful than linkage analysis in detecting genetic variants contributing to disease risk with modest effect [3]. Since the first GWAS on age-related macular degeneration (AMD) [4], there have been about 2000 GWASs reported so far, with 14609 associations showing genome-wide significance (*p*-value $\leq 5\times10^{-8}$) for 756 different diseases/traits [5].

Replication is a commonly used validation method in scientific discoveries [6, 7]. We commonly call a study used for discovery "the primary study" and a study used for validation "the replication study". In GWASs, we regard an association as a true finding with high confidence when it shows significance in both the primary and replication studies [8, 9]. Suppose the type I error rates in the primary study and the replication study are α_1 and α_2, respectively. The probability of observing more extreme statistics is below $\frac{1}{2}\alpha_1\alpha_2$ when the association doesn't exist. (We have the factor $\frac{1}{2}$ here because only the associations showing significance in the same direction in both the primary and replication studies are replicated.) Since this is a very low probability, there is a high confidence that the replicable association is a true finding.

Even when a primary association (i.e., statistically significant association in the primary study) is true, it is only replicated with a certain probability. This is because the strengths of associations are subject to random variability due to random sampling, confounding effects and measurement errors [10]. Given information from the primary study, researchers would like to know how probable it is that a primary association will be validated in the replication study. To answer this question, we need a systematic study of the behavior of primary associations in replication studies.

Recently, *Science* published an open empirical study of reproducibility in psychology, called the Reproducibility Project [11]. Researchers tried to replicate one hundred empirical studies from three top psychology journals using high-powered designs, but only thirty six of them had significant results (*p*-value$\leq$0.05). This "replication crisis" has highly stressed the importance of studying the probability of replicating a significant association in different disciplines. Currently, there is few research to quantify replication probability of a significant association. Jaffe et al. (2013) [12] gives an example to estimate the probability of replicating results in a gene-set analyses using bootstrap method.

The aim of this paper is to systematically study primary positives in the replication study setting of GWASs and to review answers to the replication question from different points of view. Since Bayesian methods can help us integrate out the uncertainty about the underlying effect of the primary association, we will mainly focus on the Bayesian view in this paper. We refer the Bayesian replication probability as the replication rate (RR). We further describe an estimation method for RR when the summary statistics of the primary study are available. We demonstrate the two applications of RR:

1. To determine the sample size of the replication study (i.e., how many individuals are needed in the replication study to achieve a certain probability of replicating primary associations?).
2. To check the consistency between the results of the primary study and those of the replication study (i.e., are the results of the replication study consistent with RR values estimated from the primary study?). We use the Hosmer-Lemeshow test [13] in this application.

The rest of this paper is organized as follows. In the next section, we will first review the answers to the replication question from the frequentist and Bayesian view. Then we will give the mathematical definitions of RR. We will also derive the relationship among the local false discovery rate (lfdr) [14], power and RR. In Section 3, we will describe how to estimate RR using the Bayesian framework with a two-component mixture prior. In Section 4, We will show the two applications of RR. In Section 5, we will describe an R-package to estimate and apply RR in GWASs. In Section 6, we will first use simulation experiments to illustrate that the estimated RR has good prediction and calibration performance when data agree with model assumptions. Then we will show the empirical results using type 2 diabetes (T2D) data from the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) consortium [15] and total cholesterol (TC) data from the Global Lipids Genetics Consortium (GLGC) [16]. The experiments also demonstrate the usefulness of RR. In Section 7, we will discuss the limitations of current modeling and estimation method, and these provide guidance for the future work. Section 8 concludes the paper.

2 Definition of RR

As an illustration, we use a *log*(*OR*) test to identify associations. Here *log*(*OR*) stands for the logarithm of the odds ratio. We can easily generalize the model to quantitative traits with simple linear regression. In Section 6, we give an example of RR estimation for GWASs with a quantitative trait.

Let's assume study *j* (*j*=1,2 denoting primary study and replication study, respectively) has $n^{(j)}$ individuals, where $n_0^{(j)}$ of them are controls and $n_1^{(j)}$ are cases. The number of SNPs is *m*. We use π_0 to denote the proportion of null SNPs, which have no association with the disease.

For each SNP, we use *A* to represent the non-effect allele and *a* to denote the effect allele. Table 1 shows a contingency table of alleles. Using the contingency table, we can estimate the logarithm of the odds ratio:

$$\hat{\mu}^{(j)} = \log n_{00}^{(j)} - \log n_{01}^{(j)} - \log n_{10}^{(j)} + \log n_{11}^{(j)}. \tag{1}$$

The true effect size μ is ordinarily unknown. Using Woolf's method [17], we can approximate the asymptotic standard error of $\hat{\mu}^{(j)}$ (denoted as $\sigma^{(j)}$) as follows:

$$\sigma^{(j)} \approx \sqrt{\frac{1}{n_{00}^{(j)}} + \frac{1}{n_{01}^{(j)}} + \frac{1}{n_{10}^{(j)}} + \frac{1}{n_{11}^{(j)}}}. \tag{2}$$

The null and alternative hypotheses are

$$\mathcal{H}_0 : \mu = 0, \text{ vs. } \mathcal{H}_1 : \mu \neq 0, \tag{3}$$

and the corresponding test statistic is $z^{(j)} = \hat{\mu}^{(j)} / \sigma^{(j)}$. Let's assume the significance levels in the two studies are α_1 and α_2, respectively.

Since we use a two-sided test in the primary study, a SNP showing an association with the disease has an absolute value of its *z*-value that is larger than $z_{\alpha_1/2}$, i.e., $|z^{(1)}| > z_{\alpha_1/2}$, where z_u is the upper *u* quantile of the standard normal distribution (0≤*u*≤0.5). For the association validated in the replication study, the *z*-value should be consistent with the *z*-value in the primary study. Thus, $z^{(2)}$ should have the same sign as $z^{(1)}$ and should also be larger than $z_{\alpha_2/2}$ in terms of absolute value, i.e., $sgn(z^{(1)})z^{(2)} > z_{\alpha_2/2}$, where the sign function reads

$$sgn(x) = \begin{cases} 1 & \text{if x} > 0 \\ 0 & \text{if x=0} \\ -1 & \text{if x} < 0 \end{cases}. \tag{4}$$

We assume the replication study is collected independently of the primary study. For a significant association in the primary study (i.e., $|z^{(1)}| > z_{\alpha_1/2}$), we want to know what is the probability of it being validated in the replication study?

From a frequentist point of view, this replication probability is the proportion of this primary association being validated in multiple independent replication studies with the same setting. If the primary association is a false positive, the replication probability is just the type I error rate $\alpha_2 / 2$. If the primary association is a true finding, then the replication probability is the power of replicating this association, which depends on the true effect size μ. The definition of replication power is

$$\beta^{(2)}(\mu) = P(sgn(z^{(1)})Z^{(2)} > z_{\alpha_2/2} \,|\, \mu, z^{(1)}). \tag{5}$$

We do not know whether the primary association is a true finding or not, and we also do not know what the true effect size is if it is a true finding. In other words, there is uncertainty in answering our major question with the frequentist replication probability.

For this reason, we will focus on the Bayesian replication probability RR. We define RR as

$$\mathrm{RR} = P(sgn(z^{(1)})Z^{(2)} > z_{\alpha_2/2} \mid z^{(1)}), \tag{6}$$

which removes the dependence of the replication probability on the underlying true status and effect size of the primary association. We can view RR as the estimator of the frequentist replication probability in terms of minimizing the Bayes risk.

We derive the relationship among the local false discovery rate of the primary study $\mathrm{lfdr}^{(1)}$, the power of the replication study $\beta^{(2)}(\mu)$ and the RR using Bayes' formula (details are in the *Supplementary Notes*):

$$\mathrm{RR} = \mathrm{lfdr}^{(1)}(\alpha_2 / 2) + (1 - \mathrm{lfdr}^{(1)})\eta^{(2)}, \tag{7}$$

where $\mathrm{lfdr}^{(1)} = P(\mathcal{H}_0 \mid z^{(1)})$, and $\eta^{(2)} = P(sgn(z^{(1)})Z^{(2)} > z_{\alpha_2/2} \mid z^{(1)}, \mathcal{H}_1) = E(\beta^{(2)}(\mu) \mid z^{(1)}, \mathcal{H}_1)$ is the Bayesian predictive power [18] of the replication study. The Bayesian predictive power $\eta^{(2)}$ averages the power $\beta^{(2)}(\mu)$ among all possible effect size values given the test statistics in the primary study.

We can regard RR as a weighted average between the true null component $\alpha_2 / 2$ and the true associated component $\eta^{(2)}$, where $\mathrm{lfdr}^{(1)}$ and 1-$\mathrm{lfdr}^{(1)}$ are the weights, respectively. Thus, we can calculate RR once $\mathrm{lfdr}^{(1)}$ and $\eta^{(2)}$ are known.

Both $\mathrm{lfdr}^{(1)}$ and $\eta^{(2)}$ are the posterior probabilities which depend on the distribution of the underlying true effect size value μ. We need to specify a prior distribution of μ for the calculation of

$\mathrm{lfdr}^{(1)}$ and $\eta^{(2)}$. In the following section, we will use a two-component mixture prior for μ to derive their calculation formulas.

3 Estimation of RR

In each study, the *log*(*OR*) estimator $\hat{\mu}^{(j)}$ is asymptotically normally distributed with a mean μ and a standard deviation $\sigma^{(j)}$, i.e.,

$$\frac{\hat{\mu}^{(j)}-\mu}{\sigma^{(j)}} \sim N(0,1). \tag{8}$$

The true effect size μ is unknown. Research on heritability decomposition [19] and effect size distribution [20] suggests that SNPs with small effect sizes occupy a larger proportion of the associated SNPs than those with large effect sizes. Hence, a natural prior for the effect size of the associated SNPs is a Gaussian prior with mean zero. Since we don't know whether an arbitrary SNP is associated or not, we use the following two-component mixture prior for all SNPs:

$$\mu \sim \pi_0 \delta_0 + (1-\pi_0) N(0,\sigma_0^2), \tag{9}$$

where δ_0 is the distribution with point mass on zero whose probability density function (pdf) is Dirac function $\delta(x)$.

With this prior, we have $Z^{(1)} \sim N(0,1+(\frac{\sigma_0}{\sigma^{(1)}})^2)$ under $\mathcal{H}_1$. Since $Z^{(1)} \sim N(0,1)$ under $\mathcal{H}_0$, we can derive the local false discovery rate of the primary study according to Bayes' rule:

$$\begin{aligned}\mathrm{lfdr}^{(1)} &= \frac{\mathrm{P}(\mathcal{H}_0)p(\mathrm{z}^{(1)} \mid \mathcal{H}_0)}{\mathrm{P}(\mathcal{H}_0)p(\mathrm{z}^{(1)} \mid \mathcal{H}_0)+\mathrm{P}(\mathcal{H}_1)p(\mathrm{z}^{(1)} \mid \mathcal{H}_1)} \\ &= \frac{\pi_0\phi(\mathrm{z}^{(1)})}{\pi_0\phi(\mathrm{z}^{(1)})+(1\text{-}\pi_0)\phi(\frac{\mathrm{z}^{(1)}}{\sqrt{1+(\sigma_0/\sigma^{(1)})^2}})},\end{aligned} \tag{10}$$

where $\phi(x)$ is the pdf of the standard normal distribution. We can regard $\mathrm{lfdr}^{(1)}$ as the proportion of null component $\pi_0\phi(z^{(1)})$ in total probability of observing $z^{(1)}$. The posterior distribution of $Z^{(2)}$ under $\mathcal{H}_1$ is $(Z^{(2)} \mid z^{(1)}, \mathcal{H}_1) \sim N\left(z^*, \sigma^{*2}\right)$, where $z^* = \lambda\hat{\mu}^{(1)} / \sigma^{(2)}$ and $\sigma^* = \sqrt{1+\lambda\left(\sigma^{(1)} / \sigma^{(2)}\right)^2}$. Here $\lambda = 1/(1+(\sigma^{(1)}/\sigma_0)^2)$ plays a shrinkage effect. Then the Bayesian predictive power of the replication study is (details in the *Supplementary Notes*)

$$\eta^{(2)} = \Phi(\frac{sgn(z^{(1)})z^* - z_{\alpha_2/2}}{\sigma^*}), \tag{11}$$

where $\Phi(x)$ is the cumulative density function (cdf) of the standard normal distribution.

When summary statistics of the primary study are available, we can approximate the asymptotic standard error of $\hat{\mu}^{(2)}$ by substituting the observed allele frequencies from the primary study into Woolf's method:

$$\sigma^{(2)} \approx \sqrt{\frac{n_0^{(1)}}{n_0^{(2)}}\left(\frac{1}{n_{00}^{(1)}} + \frac{1}{n_{01}^{(1)}}\right) + \frac{n_1^{(1)}}{n_1^{(2)}}\left(\frac{1}{n_{10}^{(1)}} + \frac{1}{n_{11}^{(1)}}\right)}. \tag{12}$$

Clearly, RR, $\mathrm{lfdr}^{(1)}$ and $\eta^{(2)}$ depend on parameters π_0 and σ_0. Since we assume all SNPs share the same distribution structure in terms of effect size in Equation (9), we can estimate the parameters with the test statistics of the primary study.

The estimation of π_0 has been addressed in the literature of FDR control from the Bayesian point of view [21]. Suppose there is a "zero assumption" that all SNPs with p-value $> \gamma$ have almost no chance to be truly associated SNPs. Let us denote the number of those SNPs as $m_+(\gamma)$. Then its expectation is

$$E(m_+(\gamma)) = m\pi_0(1-\gamma), \tag{13}$$

which introduces an π_0 estimator

$$\hat{\pi}_0 = \frac{m_+(\gamma)}{m(1-\gamma)}. \tag{14}$$

There is a tradeoff between bias and variance when choosing γ in the estimation of π_0. Storey and Tibshirani [21] proposed a procedure without tuning the parameter γ. The automated procedure will evaluate π_0 at different γ. Then, a natural cubic spline will fit to those evaluated values. We obtain the final π_0 at $\gamma = 1$ of the fitted spline.

With the estimated π_0, we estimate σ_0^2 using the method of moments. The estimator of σ_0^2 reads (see *Supplementary Notes* for details)

$$\hat{\sigma}_0^2 = \max\left(0, \left(\frac{\sum_{i=1}^{m}(z_i^{(1)})^2 - m\pi_0}{(1-\pi_0)} - m\right) / \sum_{i=1}^{m}(1/\sigma_i^{(1)})^2\right). \tag{15}$$

By plugging π_0 and σ_0^2 into Eq. (7), (10) and (11), we obtain a RR estimator. For each primary association, we can use bootstrap method to obtain the confidence interval of RR.

4 Applications of RR

We will describe the two potential applications of RR in this subsection.

4.1 Determine the sample size of the replication study

Traditional sample size determination is based on power calculation. We need to specify a minimum detectable effect size (μ_{min}) beforehand. Then, we determine the sample size such that the calculated statistical power is larger than a threshold, e.g., $\beta^{(2)}(\mu_{min}) > 80\%$. This traditional power-based sample size determination method treats the primary study and the replication study separately. The connection between the primary and replication studies is not utilized in the design. Also, the specified μ_{min} may be arbitrary, and bias may occur in the specification of μ_{min}. These factors make the determined sample size subjective.

In the design of a replication study, the main question we need to consider is how many primary associations will be replicated in the study with a given sample size? Power does not directly address this question. For example, $\beta^{(2)}(\mu_{min}) = 80\%$ doesn't mean that 80% of primary associations can be replicated. We define the expected proportion of primary associations being validated in the replication study as the global replication rate (GRR), i.e.,

$$\mathrm{GRR} = P(sgn(z_i^{(1)})Z_i^{(2)} > z_{\alpha_2/2} \,\big|\, z_i^{(1)}, i \in S), \tag{16}$$

where $S = \{i \,\|\, z_i^{(1)} |> z_{\alpha_1/2}\}$ is the index set of primary associations. We use the subscript *i* to denote the SNP index. GRR is a comprehensive measure directly addressing the question of replication. Also, the connection between the primary and replication studies is utilized in the GRR's definition. It is more natural for us to determine the sample size of the replication study based on GRR.

Clearly, there is a relationship between GRR and RR:

$$\mathrm{GRR} = \frac{1}{|S|}\sum_{i \in S} \mathrm{RR}_i, \tag{17}$$

where |*S*| is the cardinality of *S*. Please note that RR is a monotonic increasing function of $n^{(2)}$. After we set an expected replicability for primary associations, we can use the mean value of RR to determine the sample size needed in the replication study by the bisection method.

4.2 Check the consistency between the results of the primary and replication studies

After the replication study has been done, we can use RR to check the consistency between the results of the primary study and those of the replication study. Under normal situations, the results of the replication study are consistent with RR values. If inconsistency occurs, we should be alarmed, and we need to analyze the potential sources of inconsistency. These sources may be attributed to factors influencing either the primary study's or replication study's results, such as bias and measurement errors [10].

We can use the Hosmer-Lemeshow test [13] to check the consistency. The detailed steps are as follows:

1. We partition primary associations into groups according to RR. Each group has associations with approximately equal size. Let us take an example with 10 groups. The first group refers to 1/10 of the associations having the highest RR, while the second group refers to the next 1/10 of the associations having the second decile of RR, and so on.

2. We call the proportion of the replicable associations in each group the group's replication proportion (RP). We regard the mean value of RR as the group's RR.

3. We calculate the Hosmer-Lemeshow test statistic HL by comparing RP and RR in each group:

$$\mathrm{HL} = \sum_{g=1}^{G} \frac{m_g (\mathrm{RP}_g - \mathrm{RR}_g)^2}{\mathrm{RR}_g (1 - \mathrm{RR}_g)}, \tag{18}$$

   where G is the number of groups, g is the group index and m_g is the number of primary associations in group g.

4. The null hypothesis is that the results of the replication study are consistent with the RR values. We compute the p-value using the parametric bootstrap method. That is, we resample the replication status for each primary association according to RR, and then we calculate HL again. This resampling trial is repeated R times. The p-value is the proportion of resampling trials in which HL is greater than or equal to the original HL value.

5. If the p-value is smaller than the significance level, then we reject the null hypothesis. In other words, the results of the primary and the replication studies are inconsistent, and we need to analyze the potential sources of the inconsistency.

5 Software

We create an R package called RRate, available at http://bioinformatics.ust.hk/RRate.html. The package contains the following functions:

1. repRateEst(): to estimate the RR for each associations discovered from the primary study.
2. repSampleSizeRR(): to determine the sample size of the replication study for a desired GRR value.
3. HLtest(): to check the consistency between the results of the primary study and those of the replication study.

We can use the following R code (repRateEst function) to estimate the RR in a simulation data set with 10,000 SNPs and 4,000 individuals (2,000 controls and 2,000 cases) in the primary study. The sample size in the replication study is 2,000 (1,000 controls and 1,000 cases). The significance levels in two studies are 5×10^{-6} and 5×10^{-3}, respectively. We use loose significance levels here because the number of SNPs in the example data set is small. We can use the repSampleSizeRR and HLtest functions to determine the sample size of the replication study and to check the consistency between the results of the primary and replication studies.

```
library('RRate')
alpha<-5e-6              #Significance level in the primary study
alphaR<-5e-3             #Significance level in the replication study
zalpha2<-qnorm(1-alpha/2)
zalphaR2<-qnorm(1-alphaR/2)

##Load data
data('smryStats1')       #Example summary statistics from the primary study
n2.0<-2000               #Number of individuals in control group
n2.1<-2000               #Number of individuals in case group

```

```
SE2<-SEest(n2.0, n2.1, smryStats1$F_U, smryStats1$F_A) #SE in replication study

######  RR estimation  ######
RRresult<-repRateEst(log(smryStats1$OR),smryStats1$SE, SE2,zalpha2,zalphaR2,
output=T,dir='.')
RR<-RRresult$RR          #Estimated RR

#### Sample size determination ###
n1<-4000                 #Sample size of the primary study
n2_1<-repSampleSizeRR(0.8, n1, log(smryStats1$OR),smryStats1$SE,zalpha2,zalphaR2)

#### Hosmer-Lemeshow test  ####
data('smryStats2')      #Example summary statistics from the replication study
sigIdx<-(smryStats1$P<alpha)
repIdx<-(sign(smryStats1$Z[sigIdx])*smryStats2$Z[sigIdx]>zalphaR2)
groupNum<-10
HLresult<-HLtest(repIdx,RRresult$RR,g=groupNum,dir='.')
```

6 Performance and applications of the RR with simulation and real data

6.1 Simulation

We use simulation experiments to answer the following questions:

1. Can the estimated RR predict whether a primary association will be replicated or not?

2. Is the estimated RR well calibrated as the replication probability?

We simulate 5,000 controls and 5,000 cases in the primary study, and 2,500 controls and 2,500 cases in the replication study. The number of SNPs is 1×10^{6}. The effect sizes of all SNPs are generated from the following two-component distribution:

$$\mu \sim 0.95\delta_0 + 0.05N(0,0.04). \tag{19}$$

The minor allele frequencies are randomly simulated from a uniform distribution *U*(0.05,0.5), and the prevalence of the disease is set to 1%. We use $\alpha_1 = 5\times10^{-8}$ and $\alpha_2 = 5\times10^{-5}$ as significance levels in the primary study and replication study, respectively. Here we use a conservative significance level in the replication study just because we need a number of non-replicable associations to demonstrate the performance of the estimated RR. The method is applicable to any significance level.

Figure 1 shows the comparison between RR and their true values. The two scatter plots show that RR work well in terms of estimation accuracy. This kind of experiment has been run 5 times. The root mean square error of RR in Table 2 show that RR have high estimation accuracy. We also present the estimated values of parameters (π_0 and σ_0^2) in this table.

In order to see whether the estimated RR can predict the replication status well, we use RR as a score to predict whether the association can be replicated or not. We draw the receiver operator characteristic (ROC) curve (Figure 2a) using different thresholds in the prediction. A high RR value predicts that the association will be replicated. The area under the curve (AUC) is 0.858 in this simulation. This large area indicates that RR has good prediction performance as an index of replicability. In comparison, if we use the *p*-value as an index describing replicability, a low *p*-value predicts that the primary association is replicable. The AUC is 0.848, smaller than the AUC of RR. Table 3 shows the comparison of AUC in each run. The AUC of RR is larger than the AUC of *p*-value.

We use the group partition procedure (Step 1 and 2 in Subsection 4.2) to see whether the estimated RR calibrates the replication probability well. We partition the primary associations into 10 groups according to RR. Figure 2b shows a comparison between RR and RP for the 10 groups, and we can see that these two quantities agree well. The correlation between them is 0.999. This result implies RR is well calibrated as the replication probability.

We can use RR to determine the sample size needed in the replication study to achieve an expected replicability. Figure 3 plots the estimated GRR for different sample size ratios $n^{(2)} / n^{(1)}$ (from 0.5 to 1.5). For each sample size ratio, we simulate a dataset as the replication study. We call the realized proportion of primary associations being replicated the global replication proportion (GRP). We also plot GRP for different sample size ratios in Figure 3. From the figure, we can see that GRR and GRP agree well. If we want 80% of primary associations to be replicated, we need to collect about $0.5n^{(1)} = 5,000$ individuals in the replication study.

We can also use RR to check the consistency between the results of the primary study and those of the replication study. We use the Hosmer-Lemeshow test to accomplish this task. The test statistic in this simulation experiment is 14.460, and the corresponding *p*-value is 0.105. The results of the primary and replication studies are therefore consistent.

To check whether RR estimation has good performance in the phenotype with different genetic architecture, we also apply RR estimation to simulated data with different effect size distribution:

$$\mu \sim 0.95\delta_0 + 0.05N(0, 0.0064), \tag{20}$$

$$\mu \sim 0.95\delta_0 + 0.05t_{5,0.2}, \tag{21}$$

and

$$\mu \sim 0.95\delta_0 + 0.03N(0, 0.0064)+0.02N(0, 0.04), \tag{22}$$

where $t_{5,0.2}$ is a scaled *t*-distribution with degree of freedom 5 and scaling factor 0.2. In the first case, effect sizes of true associated SNPs are weak, which is a common case in a number of psychiatric disorders [22]. In the second case, effect sizes of associated SNPs follow heavy-tail distribution. In the third case, the distribution of associated SNPs' effect sizes is a mixture of weak effect and strong effect. Figure 4 shows ROC curves and the RP-RR plot for three cases. Table 4 presents corresponding AUC values in each run. In the first case, prediction performances of RR and *p*-value are similar. In other two cases, RR performs better than *p*-value in terms of prediction. Calibration performances of RR are good in all three cases. The well-calibration of RR implies that we can use RR to determine sample size of the replication study and to check the consistency between results of the primary and replication studies.

To check the performance of RR estimation when independent assumption is violated, we use GWAsimulator [23] to generate genotype data of the individuals in the general population based on haplotype distributions from the HapMap CEU samples. We simulate m=314,174 SNPs which are on Illumina HumanHap300 chip, in which 15 of them are chosen as disease loci. For each disease loci, we use multiplicative model as disease model and the relative risk for heterozygotes is 2. We simulate 5,000 controls and 5,000 cases in the primary study, and 2,500 controls and 2,500 cases in the replication study. Figure 5 shows ROC curves and the RP-RR plot. Table 5 presents AUC values in 5 runs. In this situation, RR has similar prediction performance to p-value. However, RR is well-calibrated as the replication probability.

6.2 Real data

6.2.1 T2D data from DIAGRAM

We use the public T2D dataset from DIAGRAM (http://diagram-consortium.org/) to further check RR prediction and calibration performance. We use GWAS meta-analysis (DIAGRAMv3) as the primary study. There are 56,862 individuals in the control group and 12,171 individuals in the case group. The SNP number is m=2,468,203. We use metabochip meta-analysis as the replication study. The sample size in the control group is 58,119, and the sample size in the case group is 22,669. After filtering out SNPs with p-value<0.01 in homogeneity test, there are 85,728 SNPs remaining in the replication study. We use the genome-wide significance level $\alpha_1 = 5\times10^{-8}$ in the primary study. And we use a stringent significance level $\alpha_2 = 5\times10^{-5}$ in the replication study. The reason we use this stringent threshold is that we need a number of non-replicable associations to demonstrate the performance of the estimated RR. The method is applicable to any significance level.

The estimated proportion of null hypotheses is $\hat{\pi}_0 = 0.964$, and the estimated effect size variance is $\hat{\sigma}_0^2 = 1.878 \times 10^{-3}$. There are 166 SNPs showing significant associations with T2D in the primary study and genotyped in the replication study. The GRP is 91.6%. We show the estimated RR results of these SNPs in *Supplementary Table 1* (http://bioinformatics.ust.hk/RRate.html). The estimated GRR is 90.6%, which is very close to GRP.

We draw the ROC curve for the prediction of replicability based on RR (Figure 6a). The AUC is 0.833. In comparison, if we use the *p*-value as an index predicting replicability, then the AUC is 0.762, smaller than the AUC of RR.

In order to see whether the estimated RR is well-calibrated as the replication probability, we partition all primary associations into five groups according to their RR values. Then we make a comparison between RR and RP in each group. We use five groups here instead of the 10 groups used in the simulation study because the number of primary associations is much smaller than the number of associations in the simulation experiments. Figure 6b shows the comparison between RR and RP. These two quantities agree well, with the correlation between them being 0.986. This result illustrates that RR has a good calibration performance.

We can use RR to determine the sample size of the replication study to achieve an expected number of replicable associations. We plot the estimated GRR for different sample size ratios $n^{(2)} / n^{(1)}$ (from 0.5 to 1.0) in Figure 7. If we want 80% of primary associations to be replicated, we need about $0.9n^{(1)} = 62,130$ individuals in the replication study.

We can use the Hosmer-Lemeshow test to check the consistency between the results of the primary and replication studies. The test statistic is 1.559, and the corresponding *p*-value is 0.812. The results of the primary and replication studies are therefore consistent.

6.2.2 TC data from GLGC

We also conducted experiments using published TC data from GLGC (http://csg.sph.umich.edu//abecasis/public/lipids2013/). The phenotype value measured in the study is quantitative, and we use the standardized regression coefficients as the test statistics. We use GWAS meta-analysis, comprising 94,595 individuals, as the primary study. The SNP number is *m*=1,362,710. We use metabochip study, comprising 93,982 individuals, as the replication study. After filtering out SNPs with *p*-value<0.01 in homogeneity test, there are 48,064 SNPs remaining in the replication study. The significance levels in the primary and replication studies are 5×10^{-8} and 5×10^{-5}, respectively.

The estimated proportion of null hypotheses is $\hat{\pi}_0 = 0.951$, and the estimated effect size variance is $\hat{\sigma}_0^2 = 2.346 \times 10^{-4}$. There are 631 SNPs showing statistically significant associations with TC in the primary study and genotyped in the replication study. The GRP is 88.9%. We show the estimated RR results of these SNPs in *Supplementary Table 2* (http://bioinformatics.ust.hk/RRate.html). The estimated GRR is 89.9%, which is very close to GRP.

We draw the ROC curve for the prediction of replicability based on RR (Figure 8a). The AUC is 0.871. In comparison, if we use the *p*-value as an index predicting replicability, then the AUC is 0.828, smaller than the AUC of RR.

To see whether the estimated RR has good calibration performance, we partition the primary associations into five groups according to their estimated RR values. Then we make a comparison between RR and RP in each group. Figure 8b shows the good agreement between RR and RP in the five groups. The correlation coefficient is 0.993.

We can use RR to determine the sample size of the replication study, and we plot GRR for different sample size ratios $n^{(2)} / n^{(1)}$ in Figure 9. If we want 80% of primary associations to be replicated, we need about $0.76n^{(1)} = 71,892$ individuals in the replication study.

We can use the Hosmer-Lemeshow test to check the consistency between the results of the primary and replication studies. The test statistic is 4.682, and the corresponding *p*-value is 0.29. The results of the primary and replication studies are therefore consistent.

7 Discussion

Please note that if $\mathrm{lfdr}^{(1)} > 0$, which is usually the case for a primary positive association, RR has an upper limit which is smaller than 1. According to Eq. (7),

$$\begin{aligned} \mathrm{RR} &= \mathrm{lfdr}^{(1)}(\alpha_2 / 2) + (1 - \mathrm{lfdr}^{(1)})\eta^{(2)} \\ &\le 1 - \mathrm{lfdr}^{(1)}(1 - \alpha_2 / 2), \end{aligned} \tag{23}$$

where equality is achieved if $\eta^{(2)} = 1$. The influence of the null distribution (namely $\alpha_2 / 2$) never disappears for a primary positive association with $\mathrm{lfdr}^{(1)} > 0$. The Bayesian predictive power $\eta^{(2)}$ can be increased by increasing the sample size of the replication study. In the situation of $\mathrm{lfdr}^{(1)} > 0$, no matter how many individuals participate in the replication study, the primary association will not have 100% probability of being replicated. There is also an upper bound for GRR according to Eq. (17):

$$\mathrm{GRR} \le 1 - \frac{1}{|S|} \sum_{i \in S} \mathrm{lfdr}^{(1)}{}_i (1 - \alpha_2 / 2). \tag{24}$$

If α_1 is stringent enough, $\mathrm{lfdr}^{(1)}$ is very small for each primary association. In this situation, the upper bounds of RR and GRR are close to 1.

We normally use an unbiased testing method, i.e., $\alpha_2 / 2 \le \eta^{(2)}$. Hence, we have $\mathrm{RR} \le \eta^{(2)}$ according to Equation (7). This indicates that the probability of a primary association being replicated is smaller than the Bayesian predictive power of the replication study. Normally, $\text{lfdr}^{(1)}$ is controlled by using stringent significance level α_1, and α_2 is a small value. RR and $\eta^{(2)}$ are close to each other in this situation.

At first glance, one may regard the *p*-value as a quantitative index of replicability. A statistically significant association with a lower *p*-value has a higher possibility of being replicated than an association with a higher *p*-value. The argument is that the *p*-values of associations have the same ordering as the local false discovery rates, which are the probabilities of the corresponding hypotheses being true null hypotheses given their test statistics. But a low probability of being null hypotheses does not mean a high probability of being replicated. Hence, unlike RR, the *p*-value is not directly an index of replicability.

Claiming an association to be replicated depends on the significance level of the replication study α_2. For primary associations, estimated RR values also depend on α_2. We present experiment results for $\alpha_2 = 5\times10^{-4}$ and $\alpha_2 = 5\times10^{-6}$ on our website (http://bioinformatics.ust.hk/RRate.html). From Eq. (7) and (11), RR is a monotonic increasing function of α_2. This is consistent with our intuition: the more stringent threshold we set, the less primary associations will be replicated.

RR also depends on parameters π_0 and σ_0^2 among all SNPs. From Eq. (7) and (10), RR is a monotonic decreasing function of π_0. This is because the increase of π_0 reduces the probability of each primary association being true associated one. From Eq. (7) (10) and (11), RR is a monotonic increasing function of σ_0^2. This is because the decrease of σ_0^2 increases $\text{lfdr}^{(1)}$ and decreases the power in the replication study. For diseases with weak effect sizes in associated SNPs, σ_0^2 is small and so is RR.

In some GWASs, the number of detected primary associations is small. This may be attributed to large value of π_0 and/or small value of σ_0^2. In either case with a given $n^{(2)}$, RR is small. We need large sample size in the replication study to replicate primary associations with adequate RR. Since we use the replication status of each primary association as an observation in our consistency checking method, there is no enough number of observations when only a few primary associations are detected. Hence, our consistency checking method has no enough power to detect inconsistency between results in this situation.

The accuracy of RR estimation relies on the accuracy of $\hat{\pi}_0$. Although we apply the method of Storey and Tibshirani [21] to estimate π_0, there exist other options. For example, when the "zero assumption" is violated in data or the true null distribution of test statistics does not agree with the theoretical distribution [24], it may be better to use the methods proposed by Langaas et al. [25] or Jin and Cai [26] for a reliable estimation of π_0.

Our method can be directly generalized to any tests within z-test scheme and with closed-form expression for standard error of effect size, such as log(OR) test, regression slope test and the Cochran-Armitage trend test [27]. For other tests within z-test scheme but without closed-form expression for standard error of effect size, we have problem in estimating $\sigma^{(2)}$. However, if control-to-case ratios are the same in the primary and replication studies, then we can approximate $\sigma^{(2)}$ by the central limit theorem:

$$\sigma^{(2)} \approx \sqrt{\frac{n^{(1)}}{n^{(2)}}}\sigma^{(1)}. \tag{25}$$

In this case, we can use our method to estimate RR directly.

The current model of RR is limited by the independence assumption between SNPs. In reality, correlations between SNPs, such as linkage disequilibrium, are common. An adjusted model for RR considering correlation is needed in the future.

Our current model of RR is limited to z-test scheme in single-marker test. Recent developments in sequencing technique have extended targets of association studies to both common variants and rare variants. Single-marker test is underpowered in detecting rare variants. To deal with this issue, a lot of multi-marker test and collapsing tests are proposed [28, 29]. Adjusted models for RR calculation with these testing methods are needed in the future.

8 Conclusion

In GWASs, statistically significant associations identified in a primary study need to be validated in a replication study. In this paper, we present a Bayesian framework to systematically study the behavior of those primary associations in the replication study. RR is a probabilistic measure to quantify that behavior. We describe an estimation method for RR based on the summary statistics of the primary study. We can use RR to determine the sample size of the replication study and to check the consistency between the results of the primary study and those of the replication study. We describe an R-package to estimate and apply RR in GWASs. Experiments using simulation and real data show the estimation results can accurately predict the replicability and is well calibrated. They also demonstrate the usefulness of RR.

Funding

This paper was partially supported by a theme-based research project T12-402/13N of the Hong Kong Research Grant Council (RGC).

Acknowledgement

Thanks to the Royal Academy of Engineering for a Research Exchanges with China Award to J.-H. Xue and W. Yu. We thank Prof. D. Donoho at Stanford University for insightful discussions.

References

[1] Hirschhorn JN and Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 2005;6(2):95-108.

[2] Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Reviews Genetics 2002;3(5):391-397.

[3] Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273(5281):1516-1517.

[4] Klein RJ, Zeiss C, Chew EY et al.. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308(5720):385-389.

[5] Hindorff LA, MacArthur J, Morales J et al.. A catalog of published genome-wide association studies. http://www.genome.gov/gwastudies/ (28 May 2015, date last accessed).

[6] Falk R. Replication-a step in the right direction commentary on Sohn. Theory & Psychology 1998;8(3):313-321.

[7] Blainey P, Krzywinski M, Altman N. Points of significance: replication. Nature Methods 2014;11(9):879-880.

[8] NCI-NHGRI Working Group on Replication in Association Studies. Replicating genotype–phenotype associations. Nature 2007;447(7145):655-660.

[9] Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. Statistical Science 2009;24(4):561.

[10] Ioannidis JP. Non-replication and inconsistency in the genome-wide association setting. Human Heredity 2006;64(4):203-213.

[11] Open Science Collaboration. Estimating the reproducibility of psychological science. Science 2015;349(6251):aac4716.

[12] Jaffe AE, Storey JD, Ji H et al.. Gene set bagging for estimating the probability a statistically significant result will replicate. BMC bioinformatics 2013;14(1):360.

[13] Hosmer DW, Lemesbow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics-Theory and Methods 1980;9(10):1043-1069.

[14] Efron B. Local false discovery rates. Department of Statistics, Stanford University: Technical Report 2005-20B, 2005.

[15] Morris AP, Voight BF, Teslovich TM et al.. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics 2012;44(9):981.

[16] Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. Nature Genetics 2013;45(11):1274-1283.

[17] Woolf B. On estimating the relation between blood group and disease. Ann Hum Genet 1955;19(4):251-253.

[18] Lecoutre B. Bayesian predictive procedure for designing and monitoring experiments. Luxembourg: Bayesian Methods with Applications to Science, Policy and Official Statistics, 2001,301-310.

[19] Yang J, Benyamin B, MeEvoy BP et al.. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 2010;42(7):565-569.

[20] Park JH, Wacholder S, Gail MH et al.. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature Genetics 2010;42(7):570-575.

[21] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 2003;100(16):9440-9445.

[22] Collins, AL, Sullivan, PF. Genome-wide association studies in psychiatry: what have we learned?. The British Journal of Psychiatry 2013;202(1):1-4.

[23] Li C, Li M. GWAsimulator: a rapid whole-genome simulation program. Bioinformatics 2008;24(1):140-142.

[24] Efron B. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. Journal of the American Statistical Association 2004;99:96-104.

[25] Langaas M, Lindqvist BH, Ferkingstad E. Estimating the proportion of true null hypotheses, with application to DNA microarray data. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2005;67(4):555-572.

[26] Jin J, Cai T. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. Journal of the American Statistical Association 2007;102(478):495-506.

[27] Cochran WG. Some methods for strengthening the common chi-squared tests. Biometrics 1954;10(4): 417-451.

[28] Ionita-Laza I, Lee S, Makarov V et al.. Sequence kernel association tests for the combined effect of rare and common variants. The American Journal of Human Genetics 2013;92(6):841-853.

[29] Bansal V, Libiger O, Torkamani A et al.. Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics 2010;11(11):773-785.

Key Points

1. In GWAS, positive findings often need to be validated by replication studies.

2. RR refers to the Bayesian probability of replicating a positive finding from the primary study.

3. Before collecting the replication study, we can use RR to determine the sample size of the replication study.

4. After the collection, we can use RR to check the consistency between the results of the primary study and those of the replication study.

Figures

Figure 1: The estimation method can estimate RR accurately. The x-axis is the true values of RR in the simulation study, and the y-axis is the corresponding estimated values RR. The solid line is y = x.

Figure 2: RR has good prediction and calibration performance in the simulation study. (a) We use RR and the *p*-value as scores to predict the replicated/non-replicated status in the replication study. We draw the corresponding ROC curves. The x-axis gives the false positive rate in the replicability prediction, and the y-axis gives the corresponding true positive rate. The AUC is the area under the ROC curve. RR has better prediction performance than the *p*-value. (b) We partition primary associations into 10 groups according to RR. The x-axis gives the RR of the group, which is the mean value of RR within the group. The y-axis gives the corresponding RP of the group, which is the proportion of the replicated associations in each group. The solid line is y=x. The correlation coefficient between RR and RP is 0.999. RR is well-calibrated.

Figure 3: GRR and GRP for different sample size ratios $n^{(2)}/n^{(1)}$ in the simulation experiment. We estimate GRR using summary statistics from the primary study. For each sample size ratio, we simulate a dataset as the replication study. GRP is the realized proportion of primary associations being replicated. GRR and GRP agree well in the experiment.

Figure 4: When effect sizes follow the distribution of Eq. (20), (21) and (22), the prediction and calibration performance of RR. (a) (c) (e) ROC curves of RR and *p*-value when effect sizes follow Eq. (20), (21) and (22), respectively. (b) (d) (f) RP-RR plot when effect sizes follow Eq. (20), (21) and (22), respectively.

Figure 5: When we generate genotype data in the general population based on haplotype distributions from the HapMap CEU samples, the prediction and calibration performance of RR. (a) ROC curves of RR and *p*-value. (b) RP-RR plot.

Figure 6: RR has good prediction performance and is well calibrated in T2D data from DIAGRAM. (a) We draw both the ROC curve based on RR (solid line) and the ROC curve based on the *p*-value (dashed line) in the figure. According to their AUC values, RR predicts replicability better than the *p*-value. (b) We partition primary associations into five groups according to RR. The x-axis gives the RR of the group, which is the mean value of RR within the group. The y-axis gives the corresponding RP of the group, which is the proportion of the replicated associations in each group. The solid line is *y*=*x*. RR is well calibrated.

Figure 7: GRR for different sample size ratios $n^{(2)} / n^{(1)}$ in T2D data from DIAGRAM. We use the summary statistics from the primary study to estimate GRR for different sample size ratios.

Figure 8: RR has good prediction and calibration performance in TC data from GLGC. (a) We use RR and the *p*-value to predict the replication status of primary associations, respectively. We draw their ROC curves in the figure. According to their AUC values, RR has better prediction performance than the *p*-value. (b) We partition primary associations into five groups according to RR. The x-axis gives the RR of the group, and the y-axis gives the corresponding RP of the group. The solid line is *y*=*x*. RR is well calibrated.

Figure 9: GRR for different sample size ratios $n^{(2)} / n^{(1)}$ in TC data from GLGC. We use the summary statistics from the primary study to estimate GRR for different sample size ratios.

Tables

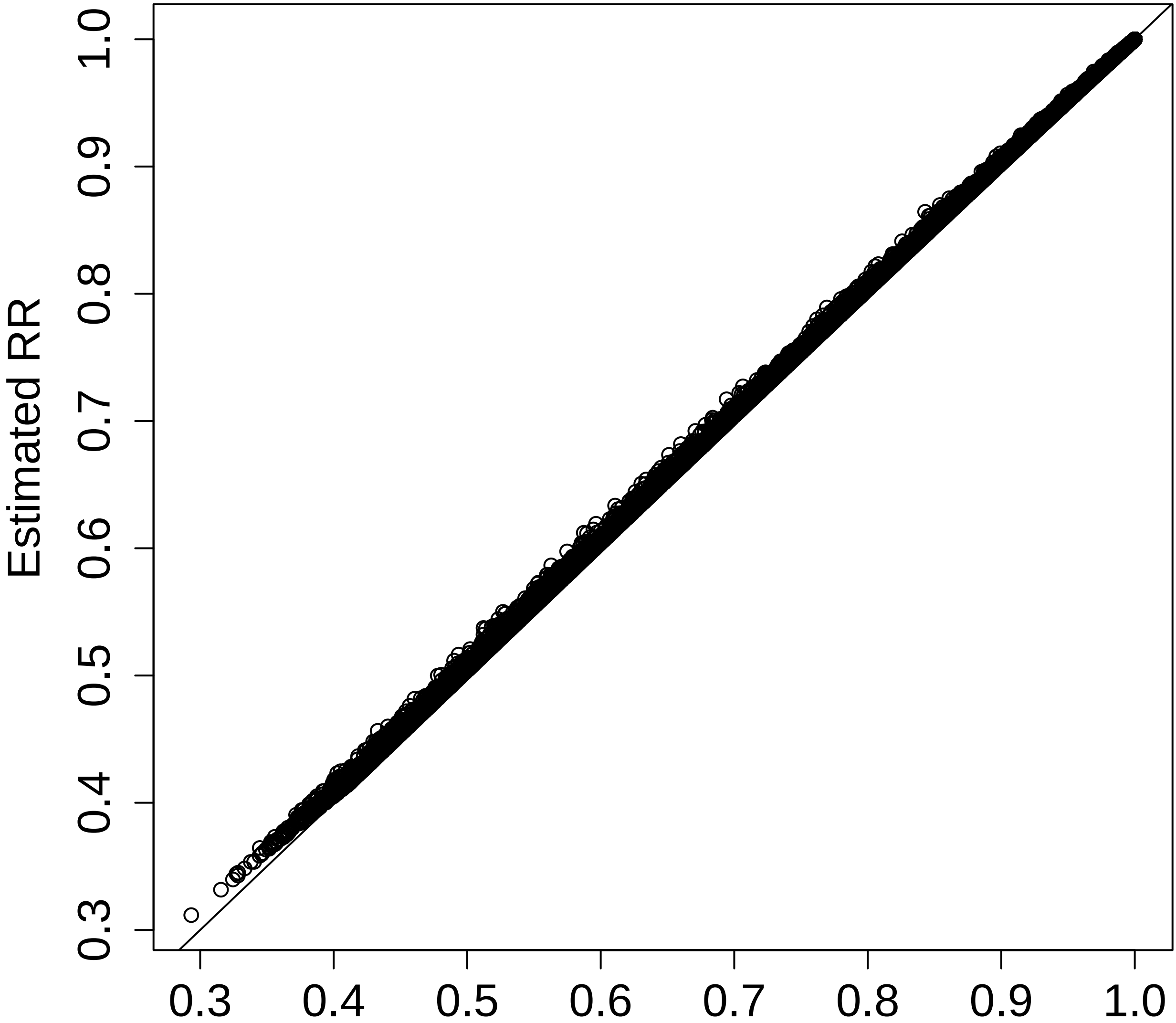
Table 1: Contingency table of one SNP in study *j*. Please see the main text for explanation of the notations.

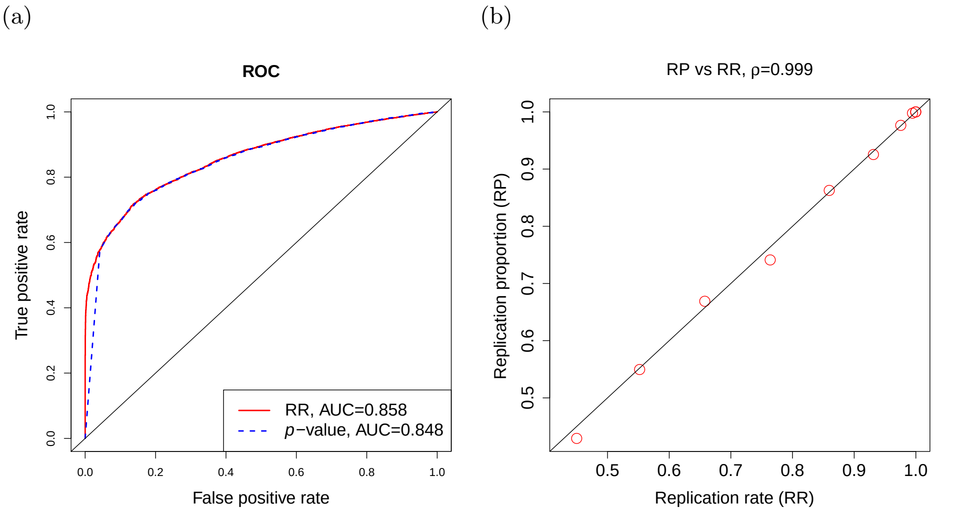
Table 2: Root mean square error of RR in simulation experiments. We also present the estimated values of parameters (π_0 and σ_0^2) in this table. The true values of π_0 and σ_0^2 are 0.95 and 0.04, respectively.

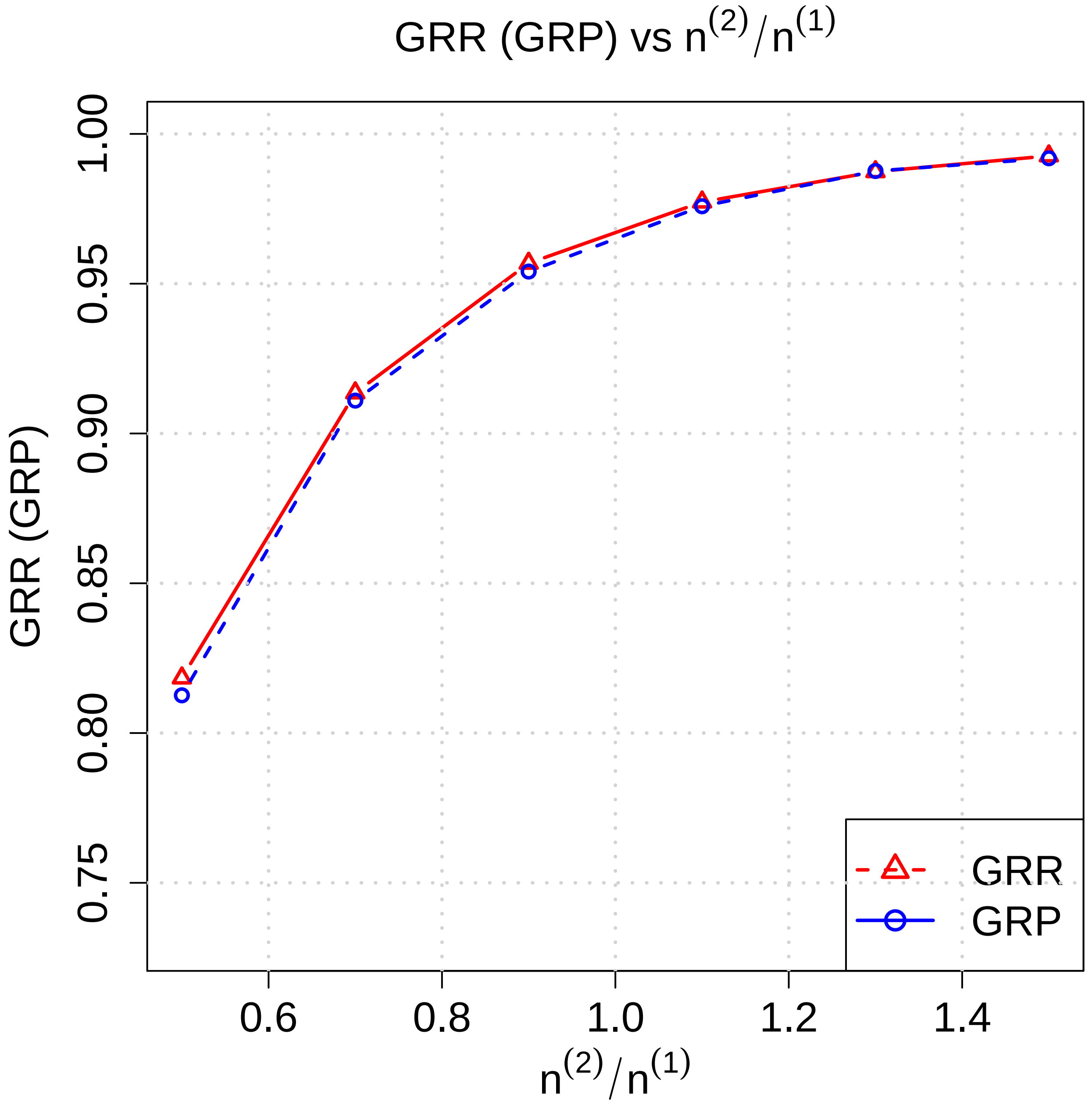
Table 3: AUC values in simulation experiments. We use RR and *p*-value as an index describing replicability, respectively.

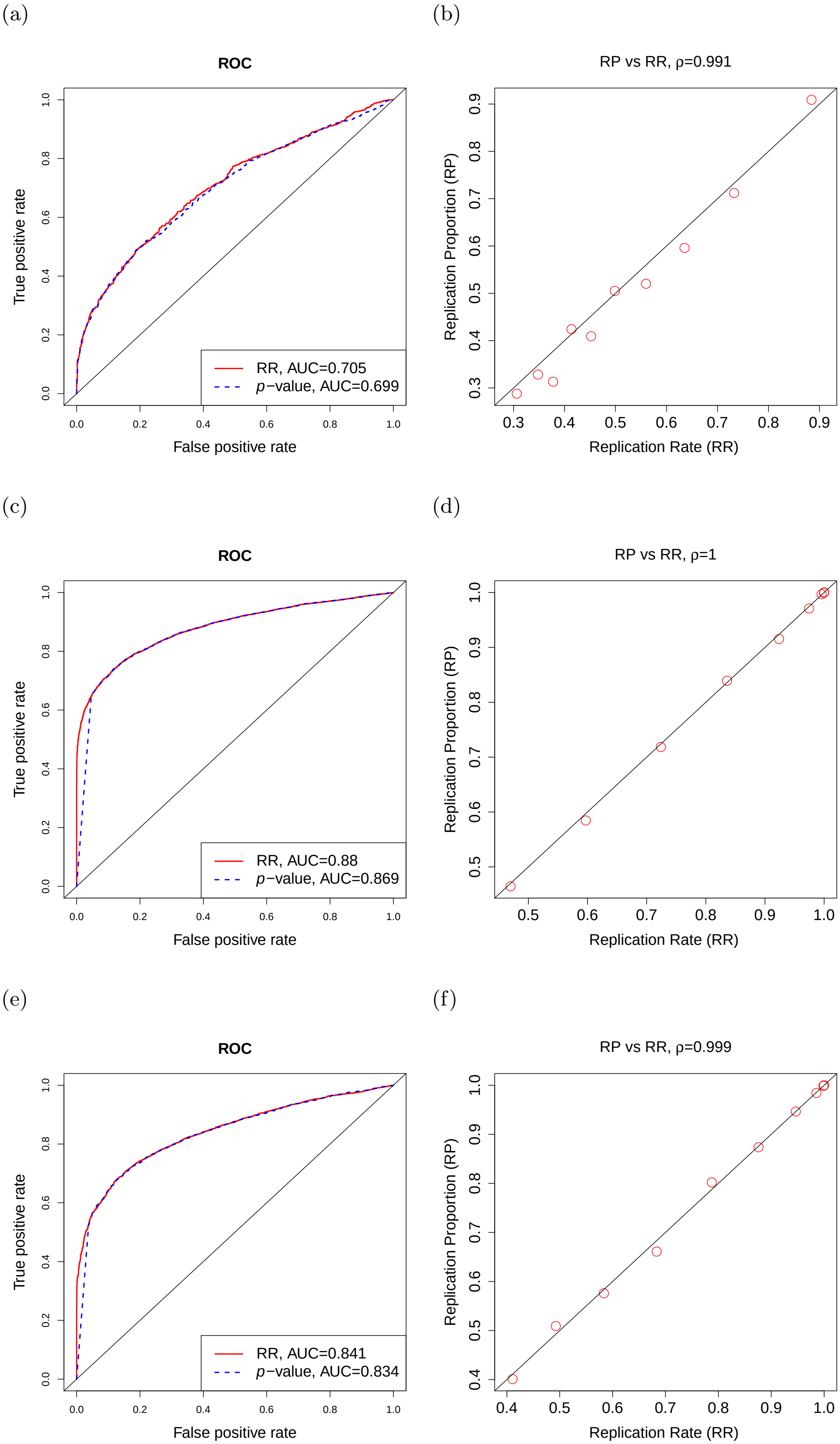
Table 4: AUC values in simulation experiments when effect sizes follow different distributions. Case 1: effect sizes follow the distribution of Eq. (20); Case 2: effect sizes follow the distribution of Eq. (21); Case 3: effect sizes follow the distribution of Eq. (22).

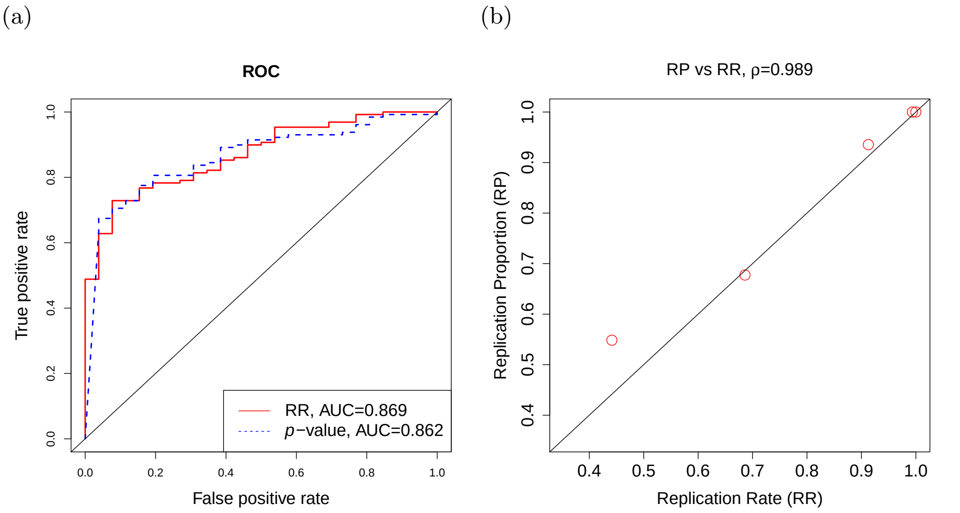
Table 5: AUC values when we generate genotype data in the general population based on haplotype distributions from the HapMap CEU samples.

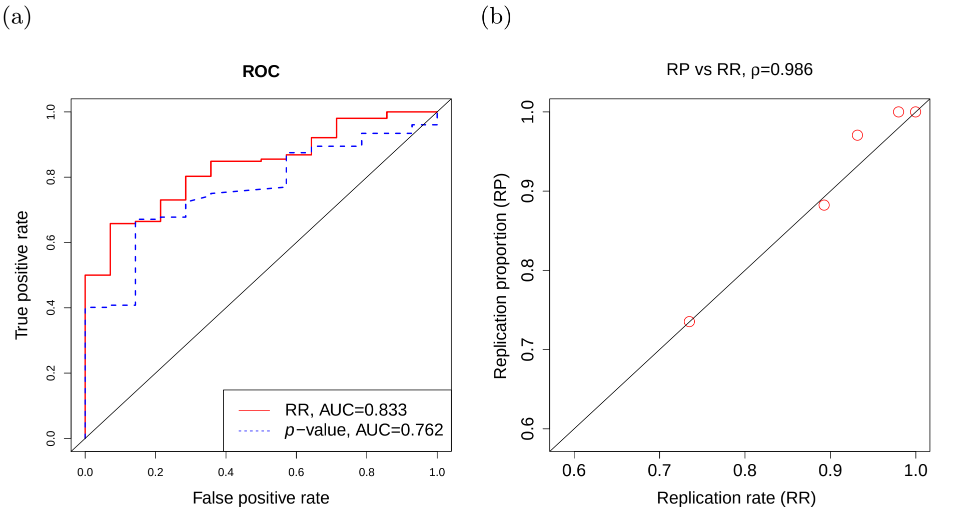

Estimated RR
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
True value of RR

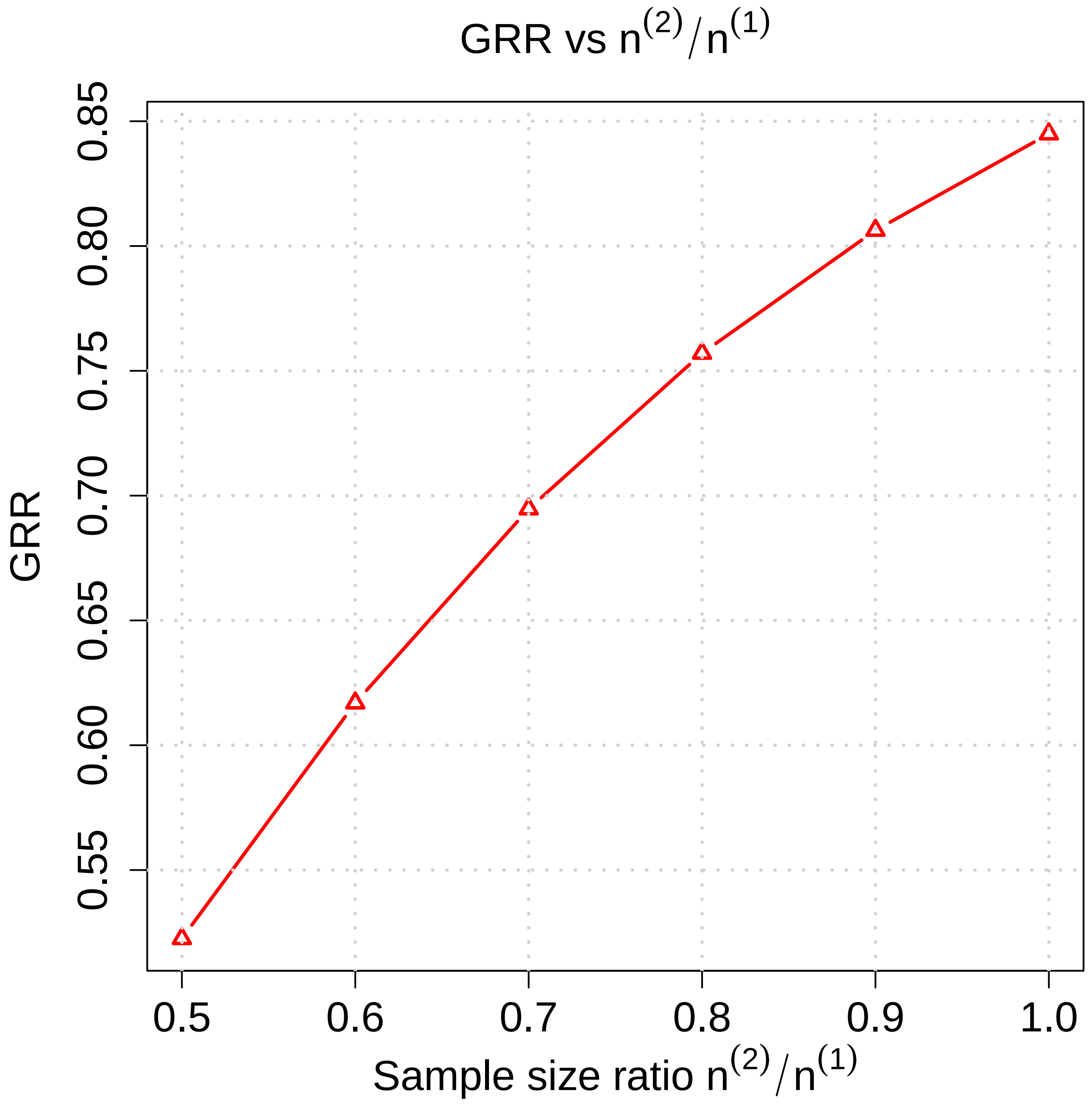

(a)
ROC
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
RR, AUC=0.858
p−value, AUC=0.848
(b)
RP vs RR, ρ=0.999
Replication proportion (RP)
0.5
0.6
0.7
0.8
0.9
1.0
Replication rate (RR)
0.5
0.6
0.7
0.8
0.9
1.0

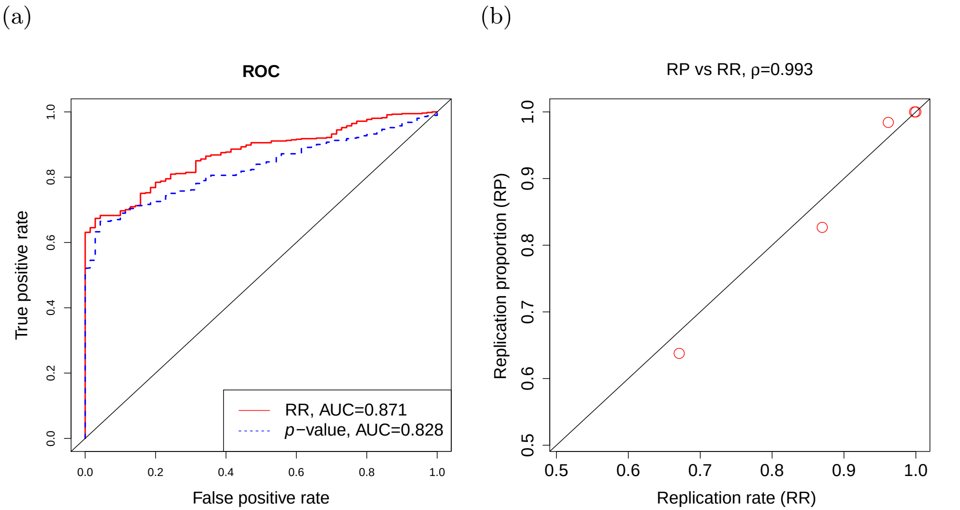

GRR (GRP) vs $n^{(2)}/n^{(1)}$
GRR (GRP)
1.00
0.95
0.90
0.85
0.80
0.75
0.6
0.8
1.0
1.2
1.4
$n^{(2)}/n^{(1)}$
GRR
GRP

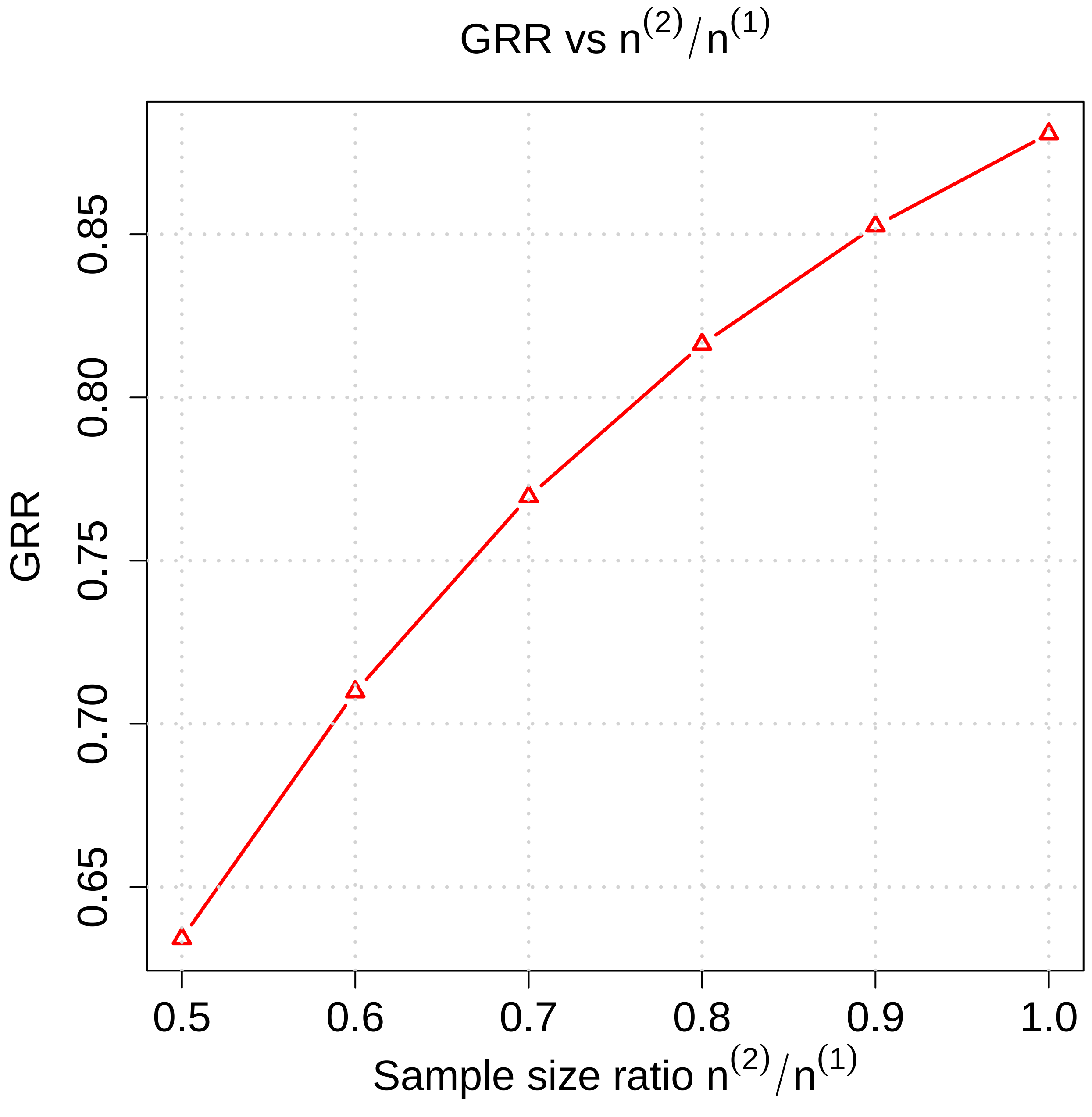

(a)
ROC
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
RR, AUC=0.705
p−value, AUC=0.699
(b)
RP vs RR, ρ=0.991
Replication Proportion (RP)
Replication Rate (RR)
0.3
0.4
0.5
0.6
0.7
0.8
0.9
(c)
ROC
True positive rate
False positive rate
RR, AUC=0.88
p−value, AUC=0.869
(d)
RP vs RR, ρ=1
Replication Proportion (RP)
Replication Rate (RR)
0.5
0.6
0.7
0.8
0.9
1.0
(e)
ROC
True positive rate
False positive rate
RR, AUC=0.841
p−value, AUC=0.834
(f)
RP vs RR, ρ=0.999
Replication Proportion (RP)
Replication Rate (RR)
0.4
0.5
0.6
0.7
0.8
0.9
1.0


(a)
ROC
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
RR, AUC=0.869
p−value, AUC=0.862
(b)
RP vs RR, ρ=0.989
Replication Proportion (RP)
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Replication Rate (RR)
0.4
0.5
0.6
0.7
0.8
0.9
1.0


(a)
ROC
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
RR, AUC=0.833
p−value, AUC=0.762
(b)
RP vs RR, ρ=0.986
Replication proportion (RP)
0.6
0.7
0.8
0.9
1.0
Replication rate (RR)
0.6
0.7
0.8
0.9
1.0


GRR vs $n^{(2)}/n^{(1)}$
GRR
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.5
0.6
0.7
0.8
0.9
1.0
Sample size ratio $n^{(2)}/n^{(1)}$


(a)
ROC
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
RR, AUC=0.871
p−value, AUC=0.828
(b)
RP vs RR, ρ=0.993
Replication proportion (RP)
0.5
0.6
0.7
0.8
0.9
1.0
Replication rate (RR)
0.5
0.6
0.7
0.8
0.9
1.0


GRR vs $n^{(2)}/n^{(1)}$
GRR
0.65
0.70
0.75
0.80
0.85
0.5
0.6
0.7
0.8
0.9
1.0
Sample size ratio $n^{(2)}/n^{(1)}$

	A	a	Total
Control	$n_{00}^{(j)}$	$n_{01}^{(j)}$	$2n_{0}^{(j)}$
Case	$n_{10}^{(j)}$	$n_{11}^{(j)}$	$2n_{1}^{(j)}$
Total	$n_{00}^{(j)} + n_{10}^{(j)}$	$n_{01}^{(j)} + n_{11}^{(j)}$	$2n^{(j)}$

Table 1. Contingency table of one SNP in study j. Please see the main text for explanation of the notations.

	RMSE of RR	$\hat{\pi}_0$	$\hat{\sigma}_0^2$
run 1	0.005	0.958	0.047
run 2	0.001	0.952	0.041
run 3	0.004	0.957	0.047
run 4	0.002	0.953	0.042
run 5	0.002	0.953	0.043

Table 2. Root mean square error (RMSE) of RR in simulation experiments. We also present the estimated values of hyperparameters (π_0 and σ_0^2) in this table. The true values of π_0 and σ_0^2 are 0.95 and 0.04, respectively.

	RR	p-value
run 1	**0.858**	0.848
run 2	**0.853**	0.843
run 3	**0.856**	0.847
run 4	**0.856**	0.846
run 5	**0.858**	0.849
Average	**0.856**	0.847

Table 3. AUC values in simulation experiments. We use RR and p-value as an index describing replicability, respectively.

	Case 1		Case 2		Case 3	
	RR	p-value	RR	p-value	RR	p-value
run 1	**0.705**	0.699	**0.880**	0.869	**0.841**	0.834
run 2	0.602	**0.603**	**0.884**	0.874	**0.846**	0.839
run 3	**0.691**	0.694	**0.884**	0.873	**0.849**	0.840
run 4	**0.643**	0.639	**0.882**	0.871	**0.845**	0.841
run 5	**0.667**	0.666	**0.883**	0.871	**0.844**	0.839
Average	**0.662**	0.660	**0.883**	0.872	**0.845**	0.839

Table 4. AUC values in simulation experiments when effect sizes follow different distributions. Case 1: effect sizes follow the distribution of Eq. (20); Case 2: effect sizes follow the distribution of Eq. (21); Case 3: effect sizes follow the distribution of Eq. (22).

	RR	p-value
run 1	**0.869**	0.862
run 2	0.822	**0.826**
run 3	**0.881**	0.873
run 4	**0.834**	0.831
run 5	**0.828**	0.826
Average	**0.847**	0.844

Table 5: AUC values when we generate genotype data in the general population based on haplotype distributions from the HapMap CEU samples.

Supplementary Notes

1 Detailed deduction of RR

The relationship between RR, $\mathrm{lfdr}^{(1)}$ and $\beta^{(2)}(\mu)$ can be derived from the law of total probability:

$$\begin{aligned}\mathrm{RR} &= P(sgn(z^{(1)})Z^{(2)} > z_{\alpha_2/2} \mid z^{(1)}) \\ &= P(sgn(z^{(1)})Z^{(2)} > z_{\alpha_2/2}, \mathcal{H}_0 \mid z^{(1)}) + P(sgn(z^{(1)})Z^{(2)} > z_{\alpha_2/2}, \mathcal{H}_1 \mid z^{(1)}) \\ &= P(\mathcal{H}_0 \mid z^{(1)})P(sgn(z^{(1)})Z^{(2)} > z_{\alpha_2/2} \mid \mathcal{H}_0, z^{(1)}) + P(\mathcal{H}_1 \mid z^{(1)})P(sgn(z^{(1)})Z^{(2)} > z_{\alpha_2/2} \mid \mathcal{H}_1, z^{(1)}) \\ &= \mathrm{lfdr}^{(1)}(\alpha_2/2) + (1-\mathrm{lfdr}^{(1)})\eta^{(2)},\end{aligned} \tag{1}$$

where

$$\begin{aligned}\eta^{(2)} &= \int_{-\infty}^{\infty} P(sgn(z^{(1)})Z^{(2)} > z_{\alpha_2/2}, \mu \mid \mathcal{H}_1, z^{(1)})d\mu \\ &= \int_{-\infty}^{\infty} P(sgn(z^{(1)})Z^{(2)} > z_{\alpha_2/2} \mid \mathcal{H}_1, \mu, z^{(1)})\, p(\mu \mid \mathcal{H}_1, z^{(1)})d\mu \\ &= \int_{-\infty}^{\infty} \beta^{(2)}(\mu)\, p(\mu \mid \mathcal{H}_1, z^{(1)})d\mu \\ &= E(\beta^{(2)}(\mu) \mid z^{(1)}, \mathcal{H}_1).\end{aligned} \tag{2}$$

2 Derivation of $\mathrm{lfdr}^{(1)}$, $\eta^{(2)}$ under a two-component mixture prior

We need the following property for multivariate Gaussian distribution to calculate $lfdr^{(1)}$ and $\eta^{(2)}$.

Property 1 *If* $\mathbf{Z} \mid \boldsymbol{\mu} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *, and* $\boldsymbol{\mu} \sim N_p(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$ *, then*

$$\mathbf{Z} \sim N_p(\boldsymbol{\mu_0}, \boldsymbol{\Sigma} + \boldsymbol{\Sigma_0}) \text{ and } \boldsymbol{\mu} \mid \mathbf{z} \sim N_p(\mathbf{W}\boldsymbol{\mu_0} + (\mathbf{I}-\mathbf{W})\mathbf{z}, (\mathbf{I}-\mathbf{W})\boldsymbol{\Sigma}) \tag{3}$$

with $\mathbf{W} = \Sigma(\Sigma_0 + \Sigma)^{-1}$.

The proof of Property 2 can be found in Chapter 2 of [1].

By using Property 2, the distribution of the test statistic $Z^{(1)}$ is

$$Z^{(1)} \sim \pi_0 N(0,1) + (1-\pi_0)N(0, 1 + (\frac{\sigma_0}{\sigma^{(1)}})^2). \tag{4}$$

Hence the local false discovery rate of the primary study can be calculated with the following:

$$\mathrm{lfdr}^{(1)} = \frac{\pi_0 \phi(z^{(1)})}{\pi_0 \phi(z^{(1)}) + (1-\pi_0)\phi(\frac{z^{(1)}}{\sqrt{1+(\sigma_0 / \sigma^{(1)})^2}})}, \tag{5}$$

where $\phi(x)$ is the pdf of the standard normal distribution.

Since $(\hat{\mu}^{(j)} \mid \mu) \sim N(\mu, (\sigma^{(j)})^2)$ and $(\mu \,\big|\, \mathcal{H}_1) \sim N(0, \sigma_0^2)$, we can obtain

$$(\mu \,\big|\, z^{(1)}, \mathcal{H}_1) \sim N(\lambda \hat{\mu}^{(1)}, \lambda (\sigma^{(1)})^2), \tag{6}$$

where $\lambda = \dfrac{1}{1+(\sigma^{(1)} / \sigma_0)^2}$ has a shrinkage effect. The posterior distribution of $Z^{(2)}$ under $\mathcal{H}_1$ reads

$$(Z^{(2)} \,\big|\, z^{(1)}, \mathcal{H}_1) \sim N\left(z^* = \lambda \frac{\hat{\mu}^{(1)}}{\sigma^{(2)}}, (\sigma^*)^2 = 1 + \lambda \left(\frac{\sigma^{(1)}}{\sigma^{(2)}} \right)^2 \right). \tag{7}$$

The Bayesian predictive power of the replication study can be calculated as follows:

$$\eta^{(2)} = \Phi(\frac{sgn(z^{(1)}) z^* - z_{\alpha_2/2}}{\sigma^*}), \tag{8}$$

where $\Phi(x)$ is the cdf of the standard normal distribution.

3 Derivation of the σ_0^2 estimator

From (4), the distribution of $Z^{(1)}$ is a two-component Gaussian mixture model. So we have

$$(Z^{(1)})^2 \sim \pi_0 \chi_1^2 + (1-\pi_0)\left(1 + (\frac{\sigma_0}{\sigma^{(1)}})^2\right) \chi_1^2, \tag{9}$$

where χ_1^2 is the χ^2 distribution with degree of freedom (df) 1. The expectation reads

$$E((Z^{(1)})^2) = \pi_0 + (1-\pi_0)(1 + (\frac{\sigma_0}{\sigma^{(1)}})^2). \tag{10}$$

For all SNPs, the following can be obtained:

$$E(\sum_{i=1}^{m}(Z_i^{(1)})^2) = m\pi_0 + (1-\pi_0)(m + \sigma_0^2 \sum_{i=1}^{m}(1/\sigma_i^{(1)})^2). \tag{11}$$

By substituting $\sum_{i=1}^{m}(z_i^{(1)})^2$ for $E(\sum_{i=1}^{m}(Z_i^{(1)})^2)$, we can get the estimator for σ_0^2:

$$\hat{\sigma}_0^2 = \left(\frac{\sum_{i=1}^{m}(z_i^{(1)})^2 - m\pi_0}{(1-\pi_0)} - m \right) / \sum_{i=1}^{m}(1/\sigma_i^{(1)})^2. \tag{12}$$

References

[1] Bishop CM. Pattern Recognition and Machine Learning. Springer, 2006.