Novel rank-based approaches for discovery and replication in genome wide association studies

Accepted to Genetics (June 2011)

This is the author's copy of the accepted article. The definitive version is available at www.genetics.org (`http://dx.doi.org/10.1534/genetics.111.130542`)

Chia-Ling Kuo[1] and Dmitri V. Zaykin[1*]

[1] National Institute of Environmental Health Sciences, National Institutes of Health

* Corresponding author's e-mail: zaykind@niehs.nih.gov

ABSTRACT

In recent years, genome-wide association studies (GWAS) have uncovered a large number of susceptibility variants. Nevertheless, GWAS findings provide only tentative evidence of association, and replication studies are required to establish their validity. Due to this uncertainty, researchers often focus on top ranking SNPs, instead of considering strict significance thresholds to guide replication efforts. The number of SNPs for replication is often determined *ad hoc*. We show how the rank-based approach can be used for sample size allocation in GWAS as well as for deciding on a number of SNPs for replication. The basis of this approach is the "ranking probability": chances that at least j true associations will rank among top u SNPs, when SNPs are sorted by P-value. By employing simple but accurate approximations for ranking probabilities, we accommodate linkage disequilibrium (LD) and evaluate consequences of ignoring LD. Further, we relate ranking probabilities to the proportion of false discoveries among top u SNPs. A study-specific proportion can be estimated from P-values, and its expected value can be predicted for study design applications.

INTRODUCTION

After a large-scale association study is completed, researchers are faced with the question of how many strongest signals to follow up on. P-values of association tests are commonly used to sort SNPs, and the most significant SNPs are considered candidates for replication. The exact number of follow-up SNPs is often determined in an informal way and may well exceed the number of SNPs deemed significantly associated. Statistical methodology can be used instead to determine that number. The rank-based approach that we develop here allows to establish a "model-based" number of follow-up SNPs, *before* conducting any association tests. In addition, this approach provides a usually lower, "data-based" number, guided by the ordered P-values that were actually observed in a study. In our approach, SNPs with prominently small P-values are followed up on, as well as any additional SNPs from the model-based calculation.

The model-based number of follow-up SNPs, u, would be determined based on an association model, to ensure a certain probability that a desired number of true associations is included. Alternatively, one can choose the number of follow-up SNPs before collecting subjects for a GWAS and find the sample size required for a desired number of true associations to be included among the chosen number of SNPs. Rank-based approach can be used in place of the usual power calculations and serves a related purpose (SATAGOPAN *et al.* 2004; GAIL *et al.* 2008; SATAGOPAN *et al.* 2002; ZAYKIN and ZHIVOTOVSKY 2005). In power calculations, the exact α (significance) threshold is predetermined, rather than the exact number of smallest P-values.

The rank-based approach has a number of attractive features.

1. SNP-specific weighting can be easily incorporated for prioritizing results of a study by specifying different prior probabilities of association for different SNPs. With this approach, SNPs in a GWAS can be weighted differentially or organized into "tiers", depending on whether they belong to a candidate pathway.

2. The approach is general in that it utilizes information contained in P-values that may originate from a variety of association tests. Parametric P-values of SNP or multilocus tests can

be used as well as permutational P-values.

3. There is only a weak dependence of ranking probabilities on LD. Probability that a true association will rank among top u SNPs quickly becomes essentially the same for correlated and independent SNPs, as u increases. This peculiar feature has been noticed previously in simulation studies (ZAYKIN and ZHIVOTOVSKY 2005) and here we investigate this phenomenon more fully.

4. It is often more appealing to researchers to think in terms of a number of SNPs rather than in terms of a significance threshold.

5. Rank-based approach has a clear connection to false discovery rate (FDR) (BENJAMINI and HOCHBERG 1995) and to Bayesian measures of association (WAKEFIELD 2007).

6. For simple models, ranking probabilities and related measures are trivially evaluated. For example, assuming 500,000 chi-square tests and noncentrality 15 for a true association, the probability that the true association will rank among the first hundred non-associated SNPs can be evaluated by a single R command (R DEVELOPMENT CORE TEAM 2007) as
```
"1-pchisq(qchisq(1-(100/(5e5+1)),df=1),df=1,ncp=15)"
```

METHODS

Ranking probabilities and FDR: Ranking probability $\mathcal{P}_{j,u}$ is the probability that at least j out of m true association P-values will have rank smaller than u in a sorted list of P-values, $\{P_{(1)},\dots P_{(k)}\}$, where k is the number of tests. A precise approximation to $\mathcal{P}_{j,u}$ under independence of P-values is given by

$$\mathcal{P}_{j,u} = F_{Y_{(j)}}\left(\frac{u-j+1}{(k-m+1)(1+1/u)}\right) \tag{1}$$

where $F_{Y_{(j)}}(\cdot)$ is the cumulative distribution function (CDF) of j-th ordered true association P-value that we denote by $Y_{(j)}$ (APPENDIX A). Equation (1) for $j = u = 1$ is the probability that one of true associations will have the smallest P-value in a study. This is also equal to power to detect at least one true association at the significance level of $1/(2L)$.

Ranking probability is closely related to the FDR. BENJAMINI and HOCHBERG (1995) suggested a method to control the expected proportion of false discoveries among discoveries (i.e. among rejections of a hypothesis) times the power:

$$\text{FDR} = E\left\{\frac{F}{T+F} \mid T+F>0\right\}\Pr(T+F>0)$$

where T, F are the numbers of true and false rejections. In association mapping scenarios, $\Pr(T+F>0)$ can decrease, as the number of tests increases, for example, if the proportion of true associations (m/k) decreases with the number of tests. Thus, the expectation part, namely, the proportion of false discoveries among discoveries, also known as conditional or positive FDR (pFDR) (STOREY 2002) may not always be controlled with this method (ZAYKIN *et al.* 2000). The pFDR is related to the false positive report probability (FPRP) (WACHOLDER *et al.* 2004) and to q-value (STOREY 2002). If the prior probability of false associations is $\Pr(H_0)$, and P-values for true effects have a common CDF $F_Y(\cdot)$, then for a fixed P-value rejection threshold α

$$\text{pFDR}(\alpha) = \text{FPRP}(\alpha) = \left[1 + \frac{1-\Pr(H_0)}{\Pr(H_0)} \times \frac{F_Y(\alpha)}{\alpha}\right]^{-1} \tag{2}$$

$F_Y(\alpha)$ is the CDF for a true association P-value. It is equal to power at the α level, i.e. the probability that a true association P-value is $\leq \alpha$. FPRP and q-value are essentially the same concepts, but the q-value approach involves estimation of $\Pr(H_0)$ and $F_Y(\cdot)$ from data. Both approaches plug in the observed P-value (p) in place of α. Thus, FPRP(p) can be interpreted as an average proportion of false positives in future experiments, provided that the α-level is readjusted to $\alpha^* = p$.

While FPRP and q-value are related to the average proportion of false positives for a fixed α-level, it is of interest to estimate the proportion of false positives at and below a particular ordered

P-value. For example, if we take u smallest P-values from a genome scan, what would be the expected proportion of false positives among $P_{(1)}, \ldots P_{(u)}$? We term this quantity the rank-FDR, rFDR(u), and express it in terms of ranking probabilities as

$$\text{rFDR}(u) = 1 - \frac{1}{u} \sum_{j=1}^{\min(u,m)} \mathcal{P}_{j,u} \tag{3}$$

In terminology of GAIL *et al.* (2008), 1–rFDR is the "proportion positive" (PP). They also introduced the "detection probability", defined as (PP×u)/m.

Equation (3) can be used for calculating the *model-based* number of follow-up SNPs, by finding the value u that controls the chosen value of rFDR. Alternatively, one can control the probability that at least j true associations are found among top u results. This is given by a single value, $\mathcal{P}_{j,u}$, and is analogous to probability of making one or more type-I errors in the usual family-wise error rate control.

The rFDR can also be expressed as an expectation using ordered P-values. Assuming a continuous test statistic, the P-value density for true associations is $f_Y(p) = dF_Y(p)/dp$. Then the posterior probability that the *observed* j-th ordered P-value (denoted by lowercase $p_{(j)}$) is a false association (i.e. posterior probability of the null hypothesis, H_0) is

$$\Pr(H_0|p_{(j)}) = \left[1 + \frac{1 - \Pr(H_0)}{\Pr(H_0)} f_Y(p_{(j)})\right]^{-1} \tag{4}$$

Note similarities and differences of Equations (2) and (4). Expectation of Equation (4), that is, the average value across many association studies, corresponds to the true proportion of $P_{(j)}$ that are false associations. The rFDR can also be expressed in terms of Equation (4):

$$\text{rFDR}(u) = E\left\{\frac{1}{u}\sum_{j=1}^{u} \Pr(H_0|p_{(j)})\right\} \tag{5}$$

A similar result cannot be obtained by averaging FPRP(p) or q-values, that is, pFDR at a fixed threshold, pFDR(α), cannot be expressed as an expectation of pFDR computed using plugged in

P-values, pFDR(p). The main distinction between the rFDR and the pFDR approaches is that the rFDR is defined for a rank, whereas pFDR is defined for a threshold.

Strictly, the right-hand side of Equation (5) is defined for a scenario where the number of true associations is not fixed, but is rather assumed to have a binomial distribution with the rate $1 - \Pr(H_0)$ and the expectation m. When m is assumed to be fixed in the computation of $\mathcal{P}_{j,u}$, there is a slight discrepancy between Equations (5) and (3) for values of m that are close to one. The distinction between random and fixed m affects the definition of $F_{Y_{(j)}}(\cdot)$. For example, when all Y's have the same distribution, then under the fixed m assumption, $F_{Y_{(1)}}(y) = 1 - (1 - F_Y(y))^m$, but with the binomial m, it is $\sum_{m=1}^{k} w_m \left[1 - (1 - F_Y(y))^m\right]$, where k is the total number of P-values and w_m are binomial probabilities of observing m true effects.

Estimation of rFDR: The model-based rFDR is the expected proportion of false discoveries, and it can be computed by Equation (3) using sample sizes, assumed number of true effects, and effect magnitudes. Its main application is planning of studies. For example, one can estimate sample size N required for the proportion of true effects to be reasonably high among the best 100 P-values. Alternatively, if N given, one can estimate the number of best results to follow up on. On the other hand, the sum of probabilities under the expectation operator in Equation (5) involves random ordered P-values that could have been observed in any particular experiment. Therefore, a *study-specific* rFDR among the u most significant P-values can be estimated as

$$\widehat{\text{rFDR}}(u) = \frac{1}{u} \sum_{j=1}^{u} \Pr(H_0 | p_{(j)}) \tag{6}$$

This requires a specification of the P-value density for true associations, $f_Y(\cdot)$, which will now be given. Asymptotic normality of a test statistic can often be assumed, and a parametric distribution such as a normal or a chi-square is often justified. Conditional on γ, the CDF of P-value for true associations can be modeled as

$$F_Y(p \mid \gamma) = 1 - G_\gamma \left(G_0^{-1}(1 - p) \right) \tag{7}$$

where $G_0(\cdot)$ and $G_\gamma(\cdot)$ denote the CDF of the test statistic under the null and the alternative hypothesis and γ is a parameter that governs power, such as noncentrality parameter of a chi-square statistic. The noncentrality is computed by substituting expected frequencies into the chi-square statistic. For example, the noncentrality of a usual chi-square test that compares allele frequencies between two groups, such as cases and controls, with sample sizes N_1 and N_2 has the noncentrality

$$\gamma = \frac{(q_1 - q_2)^2}{\left[\left(\frac{1}{2N_1} + \frac{1}{2N_2}\right)\bar{q}(1-\bar{q})\right]}$$

where q_1, q_2 are case and control allele frequencies and $\bar{q} = (N_1 q_1 + N_2 q_2)/(N_1 + N_2)$. In a logistic regression model, the noncentrality is approximately

$$\gamma = N\beta^2 q(1-q) \tag{8}$$

where β is the assumed log odds ratio, q is the population allele frequency and N is a weighted harmonic mean of N_1, N_2 with weights $q_1(1-q_1)$ and $q_2(1-q_2)$. For a common value of γ, the density $f_Y(\cdot)$ in Equation (4) can be found as follows. First, we define P-value density for a fixed γ:

$$f_Y(p \mid \gamma) \quad = \quad \frac{g_\gamma\left(G_0^{-1}(1-p)\right)}{g_0\left(G_0^{-1}(1-p)\right)}$$

where $g(\cdot)$ is the density that corresponds to the distribution $G(\cdot)$. Assuming some distribution of effect sizes among true associations, $h(\gamma)$, the density of P-values for true associations is

$$f_Y(p) \quad = \quad \int h(\gamma) f_Y(p \mid \gamma) d\gamma \tag{9}$$

Park *et al.* (2010) proposed a method to estimate both the number of true effects (m) and their distribution in large-scale association studies. They also reported tabulated frequencies of effect sizes for several diseases. In their approach, effect size is expressed as ES $= 2\beta^2 q(1-q)$; thus, given a sample size N, an empirical distribution of noncentrality can be obtained from the distribution of ES by Equation (8), i.e. $\gamma = \text{ES} \times N/2$. A probability distribution, such as gamma, can be fitted to the empirical distribution of effect sizes. This gives $h(\gamma)$, and further, when k is the total number of P-values in a study, $\Pr(H_0)$ can be taken as $1 - m/k$.

When effects sizes are defined as in PARK *et al.* (2010), ranking probabilities are related to $\widehat{\text{rFDR}}(u)$ as

$$1 - \frac{1}{u} \sum_{j=1}^{\min(u,m)} \mathcal{P}_{j,u} = E\left\{\widehat{\text{rFDR}}(u)\right\} \tag{10}$$

Alternatively, it is possible to specify a prior distribution for β (WAKEFIELD 2007), in which case the ES and the noncentrality distribution will depend on allele frequency at a particular SNP through the term $q(1-q)$. In this case, Equation (10) will no longer hold, because rankings by P-values and by posterior probabilities will not always be the same. See (WAKEFIELD 2009, 2008) for an illuminating discussion of rankings by P-values and by Bayesian measures of association.

rFDR and ranking probabilities under LD: To account for LD among L non-associated SNPs, we utilize the concept of "effective number of tests", denoted by L_e. Due to LD, L_e is generally smaller than L, and accordingly we need to consider an "effective rank", i_e as well. The effective rank parameter had been considered previously by Dudbridge in the context of false discovery rates estimation (Dudbridge, unpublished). Dudbridge and Gusnanto (DUDBRIDGE and GUSNANTO 2008) showed that the effective number of tests in a GWAS may exist and is specific to a given set of SNPs and a population. The effective number of tests in GWAS has been primarily considered in multiple testing applications, as a means to mitigate conservativeness of adjusting significance level by the number of SNPs (DUDBRIDGE and GUSNANTO 2008; MOSKVINA and SCHMIDT 2008; PE'ER *et al.* 2008; HAN *et al.* 2009; GAO *et al.* 2010). Theoretically, L_e is related to the "extremal index", θ, which arises as a parameter in the asymptotic distribution of the maximum order statistic (LEADBETTER *et al.* 1983). APPENDIX B details calculations of L_e.

Ranking probabilities that incorporate L_e can be estimated with the following approximation:

$$u_e = j + (u-j)(L_e-1)/(L-1) \quad (11)$$

$$\mathcal{P}_{j,u} \approx F_{Y_{(j)}}\left[Q_{1/2}(u_e-j+1, L_e-u_e+j)\right] \quad (12)$$

where $Q_{1/2}(u_e-j+1, L_e-u_e+j)$ is the median of Beta(u_e-j+1, L_e-u_e+j) distribution. When LD is present, but L_e cannot be estimated or guessed, we recommend use of the following approximation:

$$\mathcal{P}_{j,u} \approx F_{Y_{(j)}}\left(\frac{u-j+1}{L+1}\right) \quad (13)$$

A guess value of L_e can often be taken, because as we will show, effect of LD on ranking probabilities is generally small, and only affects probabilities at small values of u. Moreover, all three equations, (12,13,1) give similar results as u becomes larger, even when L_e/L is small. The rFDR estimates are similarly robust in the presence of LD due to their relation to ranking probabilities (Equation 10)

Simulations to evaluate accuracy of the rFDR estimation: Estimation of rFDR by Equation (6) involves specification of the effect size distribution for the assumed number of true associations. These parameters were modeled after the empirical distribution for Crohn's disease, as reported in PARK *et al.* (2010). Park and colleagues' method allows one to obtain a table of binned effect sizes (ES) with their respective frequencies. Tabulated distribution for the noncentrality for 142 susceptibility loci was obtained from the distribution of ES as $\gamma = ES \times N/2$, assuming equal sample sizes for cases and controls ($N = 2000$). The probability distribution $h(\gamma)$ in Equation (9) can be fitted to such frequency data. We assumed a gamma distribution for $h(\gamma)$ and fixed the shape parameter to be one, to allow a high proportion of very small effects. We also assumed that the distribution was truncated at the maximum observed noncentrality plus ten. The scale parameter was fitted by minimizing squared difference between quantiles of the empirical distribution and $h(\gamma)$. Alternatively, maximum likelihood or method of moments methods can be used. We assumed a two million SNP GWAS with the extent of LD modeled to provide $L_e/L = 0.4$, as has been observed in HapMap data (HAN *et al.* 2009). APPENDIX C gives details for simulation of correlated P-values. In every simulation, we recorded which ones of 15 smallest P-values were true associations. From this information, the actual proportion of true associations across simulations was computed for each of $P_{(1)}, \ldots, P_{(15)}$. These values were compared to the values $\Pr\left(H_0 \mid p_{(1)}\right), \ldots \Pr\left(H_0 \mid p_{(15)}\right)$, estimated by Equation (4) and averaged across simulations. Distribution of the actual total number of true associations among 15 smallest P-values was compared to the estimates obtained from P-values by Equation (6). We performed over 100,000 simulations to construct Table 1.

Simulations to evaluate accuracy of ranking approximations: We evaluated the accuracy of our approach using real GWAS data from a schizophrenia scan (SUAREZ *et al.* 2006). We also used simulated GWAS data with deliberately high, block-structured LD to assess the robustness of our approach in the presence of extreme LD and under conditions when the extremal index distribution (Equation B1) does not describe the asymptotic distribution of the minimum P-value. We considered the probability of capturing at least one true association and computed four types of ranking probabilities for each scenario:

1. Empirical ("gold standard") ranking probability, denoted by "True" in Tables 2–4. To calculate empirical ranking probabilities, we used a simulation approach as follows. For each of 100,000 simulations, we shuffled the affection status in the GWAS data set, and computed the trend test and its P-value for each non-associated SNP. P-values for true associations were obtained from data generated according to genetic models described below. Then we ranked P-values for both associated and non-associated SNPs, and recorded ranks of the true associations. Based on all simulations, we estimated the empirical probabilities that at least one true association will rank below the i-th false positive, $\Pr\left(\min(Y) \leq X_{(i)}\right)$. The same probabilities can also be expressed in terms of ranking below the i-th P-value rather than

below the i-th false positive (Equation A1).

2. Median-based approximate ranking probability that accounts for the LD among non-associated SNPs denoted by "L_e-based" in the tables:

$$\Pr\left(\min(Y) \leq X_{(i)}\right) \approx F_{\min(Y)}\left[Q_{1/2}(i_e, L_e - i_e + 1)\right] \tag{14}$$

   To obtain a sample of minimum P-values to estimate the effective number of tests for the calculation of the L_e-based approximation, we used a simulation approach just described. At each simulation, we recorded the minimum P-value for non-associated SNPs and used Equation (B4) to estimate L_e.

3. Median-based approximate ranking probability assuming independence of non-associated SNPs, denoted by "Median-based" in the tables:

$$\Pr\left(\min(Y) \leq X_{(i)}\right) \approx F_{\min(Y)}\left[Q_{1/2}(i, L - i + 1)\right] \tag{15}$$

4. Mean-based approximate ranking probability assuming independence of non-associated SNPs,, denoted by "Mean-based" in the tables:

$$\Pr\left(\min(Y) \leq X_{(i)}\right) \approx F_{\min(Y)}\left(\frac{i}{L+1}\right) \tag{16}$$

We conducted four simulation experiments with the first two based on the schizophrenia scan, where we retained 701,080 SNPs with a minor allele frequency of at least 0.025.

In the first experiment, we added five independent true associations with the same effect size, assuming a multiplicative model of disease risk for each SNP. In addition, we assumed multiplicative risk for joint effect of the five SNPs. We considered the Armitage trend test and retrospective sampling with 1301 cases and 1300 controls. We computed the noncentrality parameter (γ in Equation 7) by the formula given in Ahn *et. al.* (AHN *et al.* 2006). We assumed that each true association is in Hardy-Weinberg Equilibrium (HWE) in the population, with the risk allele frequency of 0.2. We used the base allele risk a=0.05 and considered the log of relative risk $\ln R$=0.206. In this model, susceptibilities for each of the five true associations are $\Pr\left(\text{case} \mid g_i\right) = a^2 R^i$, where g_i is the genotype with i copies of the risk allele. Probabilities of genotypes for the cases and for the controls were obtained by the Bayes rule. These probabilities were used to obtain the actual case and control genotypes via multinomial sampling. This setup was used to construct Table 2. Using Equation (B4), we estimated the extremal index for this experiment to be $\hat{\theta}$=$\hat{L}_e/L$=0.590.

In the second experiment, we modeled 11 SNPs in a region harboring an "unknown" causal mutation (that is, this causal SNP was removed from association analysis). The 12-SNP haplotype frequencies, including the mutation, were modeled based on the haplotype frequencies of the μ-opioid receptor (SHABALINA *et al.* 2009). These haplotype frequencies were used to create samples of haplotypes. Haplotypes were randomly paired to create diploid individuals. The fifth SNP with the minor allele frequency of 0.24 was chosen to be the causal variant. This causal SNP was contributing additively to a quantitative trait value, assuming different trait means for the two alleles, m_1=10, m_2=20, and a normally distributed random contribution with the standard deviation σ=76.72 (this value was chosen to give a similar empirical probability, $\Pr\left(\min(Y) \leq X_{(1)}\right)$, to that obtained in the first experiment). For example, the trait value for an individual i that was heterozygous at the mutation was modeled as $m_1 + m_2 + N(0, \sigma^2)$. The 12-SNP haplotype frequencies determine pairwise LD of the causal variant with the eleven markers. These eleven correlation coefficient values of LD (WEIR 1996) were 0.15, 0.17, 0.17, 0.24, 0.63, 0.78, 0.78, 0.79, 0.94, 0.95, 0.96. The causal SNP was removed from the data. Regression F-test P-values for association of the number of copies of the minor allele in an individual (0,1,2) with the trait were computed for all SNPs. The function $F_{\min(Y)}(\cdot)$ needed to compute ranking approximations, was estimated by the empirical distribution function (using "ecdf" function in R) from a sample of the minimum of P-values for the 11 SNPs. Note that to estimate the function, one only needs to obtain a single large sample of the minimum P-values. This setup was used to construct Table 3.

The third experiment was aimed to demonstrate that our approach remains valid even under unrealistically high, block-structured LD. For this experiment, we modeled correlated P-values directly by the method described in APPENDIX C. We considered a single true association, with a P-value derived from a chi-square statistic with the noncentrality parameter set to 18.42 to yield 10% power at the genome-wide significance level $0.05/L$. We assumed two million non-associated SNPs $(L = 2 \times 10^6)$, and performed 100,000 simulation replicates.

RESULTS

Table 1 gives proportions of true positives among each of 15 smallest P-values, assuming the effect size distribution and the number of disease loci estimated for Crohn's disease (PARK *et al.* 2010). There is some bias in the estimated values for low ranks, due to LD (the ratio of the effective number of tests to the actual number, L_e/L, was 0.4 in these experiments). This bias is entirely due to LD: the last two columns of the table constructed under linkage equilibrium (LE) show that the estimates are unbiased in this case. The bias can be reduced by using the following *ad hoc* correction to the prior. For i-th P-value, we modified the original prior $\Pr(H_0) = 1 - m/(L + m)$ as

$$\Pr(H_0) = 1 - (f_e/f_0)^{1/i} m/(L + m) \tag{17}$$

where $f_e = i_e/(L_e + 1)$, $f_0 = i/(L + 1)$, and i_e is the effective rank (Equation B2). The rationale behind the correction is based on the relation $\Pr\left(\min(Y) < X_{(1)}\right) = 1 - E\left\{\Pr(H_0 \mid p_{(1)})\right\}$. The minimum false association $X_{(1)}$ on the left-hand side behaves approximately as the minimum of only L_e P-values. Thus, for $P_{(1)}$, we may adjust the prior probability used to evaluate the right-hand side by the amount that is approximately equal to L/L_e. Further, we want the adjustment factor to quickly approach 1 as the rank increases. Figure 1 shows the distribution of the actual total number of true associations among 15 smallest P-values, the distribution of estimated values obtained from P-values by Equation (6), and the distribution for their difference. There is a good correspondence in the mean values for the true and the estimated distributions. The difference distribution is centered around zero, at 0.21 (-0.02 if the "corrected" prior were used instead).

There are about four true positives among 15 smallest P-values, on average. We note that in experiments with no true associations, the method still predicts about two associations on average, because the prior wrongly assumes 142 disease associations, but the probability of predicting four or more associations is less than 10%. The model-based calculation assuming independence, by using the approximation in Equation (3) shows that among the first ten smallest P-values, 31% are expected to be true positives. The true value, observed in simulations, is only slightly higher (32%). Among the first 100 P-values, only 7% are expected to be true positives; this is about 5% of the total number of true associations, 142. Using the ranking probability equation (13), we can estimate that for the probability of capturing at least five associations to be 80%, one would need to follow up on 80 smallest P-values ($\mathcal{P}_{5,80} \approx 0.8$). To capture at least ten associations, one would need to follow up on 550 strongest signals ($\mathcal{P}_{10,550} \approx 0.8$).

Rank-based calculations can be related to a more common calculation of statistical power. With the assumed parameters for the true associations, the α-level for detecting at least one true association with 90% power would have to be set at $1.1/L_e$. This is considerably higher than what would be given by multiple testing thresholds that take into account just the number of tests, such as $\alpha = 0.05/L_e$, and therefore, false positives are also expected with this approach. A comparable calculation with the rank-based approach is to find the minimum u, such that $\mathcal{P}_{1,u} \geq 0.9$, which gives $u = 3$.

The main purpose of Tables 2–4 is to illustrate accuracy of our approximations for ranking probabilities. Table 2 presents results for the 700K schizophrenia scan for five independent true associations. By comparing the empirical probability with approximate probabilities based on our approach, we studied the ranking probability that at least one of these true associations will rank among the top-i false positives (Equation A1 gives a conversion formula for ranking in terms of the top-u P-values, rather than top-i false positives). Empirical probabilities are estimates of true probabilities, since they were obtained by sorting the actual P-values, computed via the Armitage trend test that was applied at each simulation run to all 700K SNPs in LD. Empirical ranking probabilities are very similar to the approximate values in the second column, computed by the approximation that incorporates LD. Thus, our approach that incorporates the effective number of tests accounts well for LD in this data. It is apparent that ignoring the LD has very little effect on ranking probabilities for ranks as low as 10 (columns 3 and 4). Usage of either median or mean of the beta distribution (Equations 15, 16) yields similar ranking probabilities. Mean-based approximation yields values that are somewhat closer to empirical values. Similar conclusions were reached for probabilities of capturing all, rather than at least one true associations (data not shown). Results in Table 3 were obtained using the schizophrenia GWAS data as well. Here, we considered eleven correlated SNPs in a region harboring an untyped causal variant, and a continuous phenotype. We reached similar conclusions with this setup.

In addition to the results presented for the schizophrenia data set, we also evaluated our approach using two GWAS data sets for breast cancer and prostate cancer (HUNTER *et al.* 2007; YEAGER *et al.* 2007). Results for ranking probabilities obtained for these analyses were very similar to those obtained for the schizophrenia data (data not shown). Interestingly, ratio of the effective number of tests estimate over the number of SNPs, $\hat{\theta} = \hat{L}_e/L$, was very similar for the breast cancer ($\hat{\theta}$ =0.710) and the prostate cancer data sets ($\hat{\theta}$ =0.717), which both consisted of Caucasian individuals and shared most of the SNPs. This finding reaffirms conclusions by Dudbridge and Gusnanto (DUDBRIDGE and GUSNANTO 2008) who found consistency of the L_e estimates obtained for the two control data sets of the Wellcome Trust Case Control Consortium (THE WELLCOME TRUST CASE CONTROL CONSORTIUM 2007) (WTCCC) panel. The two corresponding extremal index values for the WTCCC data are similarly close: 0.58 and 0.59.

These observations suggest that once an effective number of tests estimate is obtained for a specific panel of SNPs, it can be reused for GWAS utilizing similar SNPs and samples from populations of similar ancestry. Additionally, two different designs that we used resulted in similar values of L_e. For the schizophrenia data, we obtained L_e=0.590 for the retrospective design with a case-control trend test. A similar estimate, L_e=0.610 was obtained for a prospective design with a regression test and a continuous trait.

In Table 4, we focused on the probability that a true association will rank among the top-i

highly correlated false positives. In these simulations, the effective number of tests was only 6.4% of the actual number of tests, and the behavior of the minimum P-value was incompatible with the extremal index distribution given by Equation (B1) as judged by the Kolmogorov-Smirnov test. Again, we found that the L_e-based approximation gave ranking probabilities that were very close to the empirical values across the studied range of i values. Approximations that ignore LD underestimate ranking probabilities for low ranks. Nevertheless, as i increases, these estimates quickly catch up with the empirical ranking probability. These results are reassuring in light of the fact that ranked P-values from this setup do not follow the extremal index distribution. Similar results were observed for relatively small numbers of tests (not shown).

DISCUSSION

We propose ranking-based strategies for design of large-scale association studies and for following up on best ranking associations. With our approach, one can determine an optimal number of SNPs to carry forward into a replication study, based on the sample size of the study and on the assumed model for effect sizes. This "model-based" step (Step 1) is based on calculating the proposed rFDR by Equation (3). It is reminiscent of a power calculation in that the results of association tests are not used in this step. The number of follow-up SNPs determined this way represents an expectation across multiple studies (Equation 10). In Step 2, the smallest P-values actually observed in a given study are used to estimate the study-specific rFDR (Equation 6). One would use this equation to find the largest value of u for which $\widehat{\text{rFDR}}(u)$ is at the desired level. Then u SNPs with the smallest P-values are followed up on, plus any additional ones, as determined in Step 1. The reason for following up on a potentially larger, "model-based" number, given by Step 1, is a sampling variability of study-specific estimates, illustrated in Figure 1. The averages of true and estimated values (means of the left and the middle histograms of Figure 1) correspond very well, but there is variation in both true and estimated values, as well as in their difference (given by the histogram on the right). There is a chance that none of the best results would have a high true association probability, either because none of them are true associations, or because the probability estimates happened to be incorrect. But one would not discard results of a large-scale study for the reason that too few or none of the P-values came out “non-significant”. SNP densities that are currently in use are sufficiently high to ensure coverage of truly associated regions. Even if none of the association tests in a GWAS reach statistical significance, SNPs scoring at the top still constitute putative associations.

At first glance, it may appear that unlike P-value adjustment methods, the methods that we advocate have a disadvantage in that they depend on prior parameters, namely on the effect size distribution and the number of trait loci. However, in our view, proper applications of P-value adjustments must rely on the same assumptions. Multiplicity adjustments correct the raw α-level by the number of tests, but the α-level itself must be chosen based on power calculations which are much in the same way affected by the assumed number of effects, their distribution and the sample size of the study. Two studies with different sample sizes or different number of trait loci will have a different expected proportion of true associations with P-values that are below the same α level. Relation of our approach to power can be seen more formally from our approximations: expectation for estimated posterior probability for the minimum P-value to be a true positive is equal to power at the significance level $1/(2L)$. Thus, in applications of P-value adjustment methods, the same prior parameters must be taken into account in decisions regarding an appropriate significance threshold for a study. Moreover, with our approach, it is straightforward to prioritize association results by simply specifying relatively higher prior probabilities of association to SNPs that belong to candidate pathways.

Many existing methods used for sample and marker allocations to single and multi-staged GWAS assume independence of SNPs. An analytic approach described here explicitly accommodates LD. Although simulation techniques can also be used to model LD, typically such a technique would have to utilize data from a particular GWAS. This limits applications of the simulation approach for design of future studies because the data required for simulations would have yet to be collected. A virtue of our approach is that it provides insights about the effect of LD on the ranks of true positives. In particular, it reveals consequences of the independence assumption, which have been poorly understood. Our results show that the approximation based on the effective number of tests in a GWAS (L_e) successfully accounts for the effect of LD between L SNPs. We found this method to work well even when the minimum P-value does not asymptotically follow the distribution predicted by the extremal index theory. This observation is explained by the fact that validity of our approach depends only on similarity of medians of the theoretical and the actual distributions, rather than on similarity of entire distributions. In fact, L_e does not need to exist when interpreted in the conventional sense as a parameter governing behavior of the smallest P-value. The $\theta = L_e/L$ values appear to be very similar for similar panels of SNPs and samples from populations of similar ancestry. Our estimates of θ for two different samples genotyped on the same platform are within one percent of each other. Similarly, there is only one percent difference between θ estimates for the two control samples from the WTCCC data (DUDBRIDGE and GUSNANTO 2008). Further, we showed theoretically and confirmed via applications to GWAS data that as the rank value increases, the effect of LD on the ranking probability quickly becomes negligible. This reaffirms and generalizes our earlier observation from simulation studies that ranks of true associations appear to be little affected by LD (ZAYKIN and ZHIVOTOVSKY 2005). This surprising property holds even when the extent of LD is unrealistically high. In a GWAS with L SNPs, LD for moderate values of ranks can be ignored as long as $1/L$ is much smaller than L_e/L. In practical applications, ranking probability may be set to some high value $\mathcal{P}$, such as 0.90, and one would then compute the value of rank i such that the ranking probability is greater or equal to $\mathcal{P}$. Unless the resulting value of i is small, ranking probabilities computed using the independence assumption yield values that are very close to exact values. Even for low values of ranks, bias due to usage of independence assumption is small.

The rank-based approach can be used for determination of sample size while planning a study. Just as with a power calculation, one would assume an effect size and the number of tests. Instead of determining a sample size N, needed for a true association P-value to be as small as α/L with 90% probability (as in a power calculation), one could determine N such that at least j true association

P-values will end up among a specified number (u) of best results with 90% probability. This gives the ranking probability, $\mathcal{P}_{j,u}$; alternatively, rFDR can be controlled by the sum of $\mathcal{P}_{j,u}$ (Equation 3).

In summary, our ranking approach provides a simple and intuitively appealing framework for planing and analysis of large-scale association studies. This approach is broadly applicable due to its robustness in the presence of extreme and heterogeneous LD. Although our approximations are mainly developed assuming a large number of SNPs, they work well when the number of SNPs is small, as found in candidate gene studies. These features make our approach well suited for studies with different marker densities and LD patterns, including studies utilizing next-generation sequencing data. For practical use, we provide software that allows one to estimate ranking probabilities, number of SNPs that would contain true associations with a specified probability, as well as to plan discovery and replication stages in GWAS.

URLs

Software implementing the methods described here is available at the NIEHS website, (`http://www.niehs.nih.gov/research/atniehs/labs/bb/staff/zaykin/index.cfm`), or by request to the authors.

ACKNOWLEDGMENTS

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. We benefited from discussion with Shyamal Peddada and David Umbach and from useful comments by Gary Churchill and two anonymous reviewers. Funding support for the Genome-Wide Association of Schizophrenia Study was provided by the National Institute of Mental Health (R01 MH67257, R01 MH59588, R01 MH59571, R01 MH59565, R01 MH59587, R01 MH60870, R01 MH59566, R01 MH59586, R01 MH61675, R01 MH60879, R01 MH81800, U01 MH46276, U01 MH46289 U01 MH46318, U01 MH79469, and U01 MH79470) and the genotyping of samples was provided through the Genetic Association Information Network

(GAIN). The datasets used for the analyses described in this manuscript were obtained from the database of Genotypes and Phenotypes (dbGaP) found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession numbers phs000207.v1.p1.c1, phs000021.v2.p1, phs000147.v1.p1.c1. Samples and associated phenotype data for the Genome-Wide Association of Schizophrenia Study were provided by the Molecular Genetics of Schizophrenia Collaboration (PI: Pablo V. Gejman, Evanston Northwestern Healthcare (ENH) and Northwestern University, Evanston, IL, USA).

APPENDIX A

Approximations for ranking probabilities: We consider a set of true association P-values, $\{Y_1, \ldots, Y_m\}$, and a set of false association P-values, $\{X_1, \ldots, X_L\}$, that are ordered together. Ranking probability as defined in METHODS is equivalent to the probability that the j-th true association P-value will rank below the P-value with the rank $u \leq L$, i.e. $\Pr\left(Y_{(j)} \leq P_{(u)}\right)$. We can write the ranking probability in a mathematically convenient way that separates true and false associations:

$$\mathcal{P}_{j,u} = \Pr\left(Y_{(j)} \leq P_{(u)}\right) = \Pr\left(Y_{(j)} \leq X_{(u-j+1)}\right) \tag{A1}$$

A random i-th ordered P-value among non-associated SNPs is denoted by $X_{(i)}$. Its cumulative and probability distribution functions (CDF and PDF) are denoted by $F_{X_{(i)}}(\cdot)$ and $f_{X_{(i)}}(\cdot)$, respectively. Similarly, $F_Y(\cdot)$ and $f_Y(\cdot)$ are the CDF and the PDF for a true association.

When Y's are independent and have the same distribution, the CDF of Y is given by Equation (7) and the CDF of the j-th ordered Y has a standard form, $F_{Y_{(j)}} = 1 - \text{Binom}(j-1;\, m, F_Y(p))$, where Binom denotes the binomial CDF evaluated at $j-1$ successes in m trials with the success probability $F_Y(p)$. Otherwise, the distribution $F_{Y_{(j)}}$ can be estimated by the empirical CDF from a sample of P-values obtained under a suitable association model. For O simulation samples, a $O \times m$ matrix of P-values ($\mathbf{P}$) can be obtained. After the rows of $\mathbf{P}$ are sorted, $\hat{F}_{Y_{(j)}}$ is given by the empirical CDF, e.g. by "ecdf" function of R, applied to the j-th column of the matrix. If $Y_1 \ldots Y_m$ are independent, but different effect sizes are assumed, the corresponding P-values can be sampled directly, provided the distribution of the test statistic is standard, such as chi-square: $P_i = 1 - G_0\left(G_{\gamma_i}^{-1}(U)\right)$, $i = 1, \ldots, m$, where U is a random number from the Uniform(0,1) distribution. This formula can be obtained by inverting Equation (7).

We start with an independence assumption, namely that non-associated SNPs are themselves independent. Considering a single true association for now, probability that the P-value for a truly associated SNP will rank among i smallest null P-values is given by

$$\begin{aligned}\Pr\left(Y \le X_{(i)}\right) &= 1 - \int_0^1 \int_0^y f_{X_{(i)}}(x) f_Y(y) dx dy \\ &= 1 - \int_0^1 F_{X_{(i)}}(y) f_Y(y) dy \\ &= 1 - E_Y\left\{F_{X_{(i)}}(Y)\right\} \qquad \text{(A2)}\end{aligned}$$

where $F_{X_{(i)}}$ is the CDF of Beta$(i, L - i + 1)$, if L non-associated P-values are independent and continuous. When $L = 1$, this probability equals one minus the expectation of Y, μ_Y, which was termed "expected P-value" and advocated as a sensible alternative to power calculation by Sackrowitz and Samuel-Cahn (SACKROWITZ and SAMUEL-CAHN 1999). Although it is possible to evaluate Equation (A2) by numerical integration or by simulation, a useful and precise approximation is obtained as follows. First, by means of changing the order of integration, we can rewrite Equation (A2) as

$$\Pr\left(Y \le X_{(i)}\right) = E_X\left\{F_Y\left[F_{X_{(i)}}^{-1}(1 - X)\right]\right\} \qquad \text{(A3)}$$

This operation replaces the random variable Y on the right-hand side of Equation (A2) with the uniformly distributed random variable X. By using the first degree Taylor series approximation about $E(X)$, we obtain

$$\begin{aligned}\Pr\left(Y \leq X_{(i)}\right) &\approx F_Y\left[F^{-1}_{X_{(i)}}(1/2)\right] && \text{(A4)}\\ &= F_Y\left[Q_{1/2}(i, L-i+1)\right] && \text{(A5)}\\ &\approx F_Y\left[\frac{i}{(L+1)(1+1/i)}\right] && \text{(A6)}\end{aligned}$$

where $Q_{1/2}$ is the median of Beta$(i, L-i+1)$.

APPENDIX B

Approximations for ranking probabilities under LD: For a variety of stationary stochastic processes with distance-decaying dependencies, the distribution of the maximum of a large number of observations $\{T_i\}$ is $\Pr\left(\max\left(T\right) \leq t\right) = \Pr(T \leq t)^{\theta L}$. The independence is a special case with $\theta = 1$. Considering the minimum of P-values, $\{X_i\}$, we can let $T_i = 1 - X_i$, and $\max(T) = 1 - \min(X)$. Then the asymptotic distribution of $\min(X)$ under no association is

$$\Pr\left(\min(X) < x\right) = 1 - (1-x)^{\theta L} \tag{B1}$$

which is the CDF of Beta$(1, \theta L)$. Thus, the effective number of tests is $L_e = \theta L$. The effective rank is obtained by scaling the original rank i between 1 and L_e:

$$i_e = 1 + (i-1)(L_e - 1)/(L-1) \tag{B2}$$

The ranking probability is then obtained as

$$\Pr\left(Y \leq X_{(i)}\right) \approx F_Y\left[Q_{1/2}(i_e, L_e - i_e + 1)\right] \tag{B3}$$

There is a concern that assumptions of the extremal index theory, such as stationarity, may not be satisfied at high SNP densities. Indeed, Dudbridge and Koeleman found that at HapMap densities, fit to a beta distribution may become inadequate (DUDBRIDGE and KOELEMAN 2004). However, our approach does not require that the minimum P-value should follow a beta distribution: only its median needs to be approximately equal to that of Beta$(1, \theta L)$ distribution. Our requirement is much weaker than the distributional assumption, and we will verify that our approach works well even in those cases where the effective number of tests does not exist in the sense as defined by the asymptotic distribution in Equation (B1). Dudbridge and Gusnanto (DUDBRIDGE and GUSNANTO 2008) gave the method of moments estimator for L_e as $(1 - \bar{p})/\bar{p}$ where $\bar{p}$ is the sample average of minimum P-values under the null hypothesis. It is also straightforward to derive the maximum-likelihood estimator (MLE) as

$$\hat{L}_e = -k / \sum_{i=1}^{k} \ln(1 - p_i) \tag{B4}$$

where $\{p_i\}$ is a sample of minimum P-values. In practice, one would repeat the following procedure k times to obtain that sample: permute the affection status of individuals, perform an association test for all SNPs, and record the minimum P-value. Usage of the effective number of tests approach has an advantage in that once an estimate of $\hat{\theta} = \hat{L}_e/L$ is obtained, it can be reused for computing ranking probabilities for a different data set that utilizes a similar set of SNPs and a sample from a population of similar ancestry. It is also possible to fit a two-parameter beta distribution and obtain maximum likelihood estimates (MLE) for the two parameters, as suggested by Dudbridge and Gusnanto (DUDBRIDGE and GUSNANTO 2008). These estimates can be used to model the distribution $F^{-1}_{X_{(i)}}$ in Equation (A4). Ranking probabilities evaluated by this approach according to our simulation results were very close to those based on L_e (results not shown). Based on the form of Equation (A5), we expect that the independence assumption may work well even though in practice P-values are in fact dependent, due to LD. The reason for this is the similarity of medians $Q_{1/2}(i, L - i + 1)$ and $Q_{1/2}(i_e, L_e - i_e + 1)$ for a moderate i and a large L (cf Equation (A5)). As

i increases, i_e approaches θi, and the median of an intermediate order statistic approaches its mean (WATTS *et al.* 1982). Thus, assuming that L is large and $1/L$ is much smaller than θ,

$$\begin{aligned} Q_{1/2}(i_e, L_e - i_e + 1) &\approx \frac{i_e}{\theta L + 1} \approx \frac{\theta i}{\theta L + 1} \approx \frac{i}{L+1} \\ &\approx Q_{1/2}(i, L - i + 1) \end{aligned} \tag{B5}$$

We also expect that a simpler, mean-based approximation is adequate for intermediate values of i:

$$\begin{aligned} \Pr\left(Y \leq X_{(i)}\right) &\approx F_Y\left(\frac{i_e}{L_e + 1}\right) && \text{(B6)} \\ &\approx F_Y\left(\frac{i}{L+1}\right) && \text{(B7)} \end{aligned}$$

One might be interested in the number of SNPs required to be genotyped in a follow-up study to ensure for the ranking probability to reach some high value, such as $\mathcal{P} \times 100\% = 99\%$. In other words, one might be interested in determining the index i in Equation (A5) for the probability to be at least $\mathcal{P} = 0.99$. A simple way to do that is to iterate i from one up to a value that gives the required probability. The mean-based approximation also allows one to determine i by solving Equation (B6). It follows that $i_e = \left\lceil (L_e + 1) \, F_Y^{-1}(\mathcal{P}) \right\rceil$, and from Equation (B2) $i = 1 + (i_e - 1)(L - 1)/(L_e - 1)$.

APPENDIX C

Simulation of correlated P-values: For each simulation replicate, we sampled P-values for the true associations by the inverse of Equation (7), as $P = 1 - G_0\left(G_\gamma^{-1}(U)\right)$, where U is a random number from the Uniform(0,1) distribution. We assumed the one degree of freedom chi-square statistic. P-values for non-associated SNPs were sampled from the zero-mean Ornstein-Uhlenbeck diffusion process (OUP). The OUP diffusion had been employed previously to address issues of multiple testing in genome scans (LANDER and BOTSTEIN 1989; LANDER and KRUGLYAK 1995;

ZAYKIN and ZHIVOTOVSKY 2005). Let us denote a normal score for the i-th SNP by S_i, i=1, 2, ..., $2{\times}10^6$. We simulated S_i for $i > 1$ by $S_i = S_{i-1}e^{-\lambda\Delta_i} + \sigma\sqrt{(1-e^{-2\lambda\Delta_i})/(2\lambda)}\,\epsilon_i$, where ϵ_i is a sample from the standard normal distribution, $N(0,1)$. This formula follows from considering the conditional distribution $(S_i \mid S_{i-1})$, which is normal with the mean $S_{i-1}e^{-\lambda\Delta_i}$ and the variance $\frac{\sigma^2}{2\lambda}\left(1-e^{-2\lambda\Delta_i}\right)$ (GILLESPIE 1996). After $\{S_i\}$ were simulated, we converted them to P-values using the unconditional limiting normal CDF of S_i, $N(0, \sigma^2/2\lambda)$. The value S_1 was simulated as a draw from this distribution. For simulations with extreme, block-patterned correlation we used following parameters. We used $\lambda = 0.5$ and $\sigma^2 = 100$. Each non-associated SNP had a chance of 0.7 to be in a LD block. To model block-structured correlation, the location increment parameter, Δ_i, assumed one of two values, 0.033, with probability 0.3, or 0.001, with probability 0.7. This model resulted in extremely correlated P-values, with $L_e/L = 0.064$. On average, autocorrelation decayed from 0.99 for neighboring SNPs to 0.5 for SNPs 187.5 Kb apart. Between the blocks, the autocorrelation decayed to 0.1 in 205.5 kB. Within the blocks, there was very little decay of correlation: to 0.93 in a 200 kB block. To model studies with the HapMap ratio of L_e/L=0.4, we changed the two Δ_i values to 0.1 and 0.07.

LITERATURE CITED

AHN, K., C. HAYNES, W. KIM, R. FLEUR, D. GORDON, *et al.*, 2006 The effects of SNP genotyping errors on the power of the Cochran-Armitage linear trend test for case/control association studies. Ann Hum Genet **71**: 249–261.

BENJAMINI, Y., and Y. HOCHBERG, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Stat Soc B **57**: 289–300.

DUDBRIDGE, F., and A. GUSNANTO, 2008 Estimation of significance thresholds for genomewide association scans. Genet Epidemiol **32**: 227–234.

DUDBRIDGE, F., and B. KOELEMAN, 2004 Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. Am J Hum Genet **75**: 424–435.

GAIL, M. H., R. M. PFEIFFER, W. WHEELER, and D. PEE, 2008 Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. Biostatistics (Oxford, England) **9**: 201–215.

GAO, X., L. BECKER, D. BECKER, J. STARMER, and M. PROVINCE, 2010 Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol **34**: 100–105.

GILLESPIE, D., 1996 Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral. Physical Review E **54**: 2084–2091.

HAN, B., H. KANG, and E. ESKIN, 2009 Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. PLoS Genetics **5**: 1–13.

HUNTER, D., P. KRAFT, K. JACOBS, D. COX, M. YEAGER, *et al.*, 2007 A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet **39**: 870–874.

LANDER, E., and D. BOTSTEIN, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185–199.

LANDER, E., and L. KRUGLYAK, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet **11**: 241–7.

LEADBETTER, M., G. LINDGREN, and H. ROOTZÉN, 1983 *Extremes and related properties of random sequences and processes*. Springer Verlag, Berlin.

MOSKVINA, V., and K. SCHMIDT, 2008 On multiple-testing correction in genome-wide association studies. Genet Epidemiol **32**: 567–573.

PARK, J., S. WACHOLDER, M. GAIL, U. PETERS, K. JACOBS, *et al.*, 2010 Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet **42**: 570–575.

PE'ER, I., R. YELENSKY, D. ALTSHULER, and M. DALY, 2008 Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol **32**: 381–385.

R DEVELOPMENT CORE TEAM, 2007 *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

SACKROWITZ, H., and E. SAMUEL-CAHN, 1999 P values as random variables – expected P values. American Statistician **53**: 326–331.

SATAGOPAN, J., E. VENKATRAMAN, and C. BEGG, 2004 Two-Stage Designs for Gene–Disease Association Studies with Sample Size Constraints. Biometrics **60**: 589–597.

SATAGOPAN, J., D. VERBEL, E. VENKATRAMAN, K. OFFIT, and C. BEGG, 2002 Two-stage designs for gene-disease association studies. Biometrics **58**: 163–170.

SHABALINA, S. A., D. V. ZAYKIN, P. GRIS, A. Y. OGURTSOV, J. GAUTHIER, *et al.*, 2009 Expansion of the human mu-opioid receptor gene architecture: novel functional variants. Hum Mol Genet **18**: 1037–51.

STOREY, J., 2002 A direct approach to false discovery rates. J Royal Stat Soc B **64**: 479–498.

SUAREZ, B., J. DUAN, A. SANDERS, A. HINRICHS, C. JIN, *et al.*, 2006 Genomewide linkage scan of 409 European-ancestry and African American families with schizophrenia: suggestive evidence of linkage at 8p23. 3-p21. 2 and 11p13. 1-q14. 1 in the combined sample. Am J Hum Genet **78**: 315–333.

THE WELLCOME TRUST CASE CONTROL CONSORTIUM, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447**: 661–678.

WACHOLDER, S., S. CHANOCK, M. GARCIA-CLOSAS, L. EL GHORMLI, and N. ROTHMAN, 2004 Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J Natl Canc Inst **96**: 434–442.

WAKEFIELD, J., 2007 A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet **81**: 208–227.

WAKEFIELD, J., 2008 Reporting and interpretation in genome-wide association studies. Int J Epidemiol **37**: 641–53.

WAKEFIELD, J., 2009 Bayes factors for genome-wide association studies: comparison with P-values. Genet Epidemiol **33**: 79–86.

WATTS, V., H. ROOTZEN, and M. LEADBETTER, 1982 On limiting distributions of intermediate order statistics from stationary sequences. The Annals of Probability **10**: 653–662.

WEIR, B. S., 1996 *Genetic data analysis II*. Sinauer Associates, Sunderland, Mass.

YEAGER, M., N. ORR, R. HAYES, K. JACOBS, P. KRAFT, *et al.*, 2007 Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. Nat Genet **39**: 645–649.

ZAYKIN, D., S. YOUNG, and P. WESTFALL, 2000 Using the false discovery rate approach in the genetic dissection of complex traits: a response to Weller et al. Genetics **154**: 1917–1918.

ZAYKIN, D. V., and L. A. ZHIVOTOVSKY, 2005 Ranks of genuine associations in whole-genome scans. Genetics **171**: 813–823.

TABLES

TABLE 1
Probability that u-th sorted P-value is a true association in a two-million simulated SNP scan modeled after the effect size distribution for Crohn's disease

	$\Pr\left(H_A \mid p_{(u)}\right)$ under LD			$\Pr\left(H_A \mid p_{(u)}\right)$ under LE	
Rank (u)	True	Estimated	Estimated*	True	Estimated
1	0.821	0.716	0.829	0.738	0.739
2	0.601	0.491	0.544	0.527	0.526
3	0.424	0.357	0.383	0.390	0.389
4	0.310	0.281	0.296	0.303	0.303
5	0.239	0.236	0.244	0.247	0.248
6	0.196	0.205	0.210	0.210	0.211
7	0.170	0.182	0.186	0.184	0.184
8	0.151	0.165	0.168	0.164	0.164
9	0.136	0.151	0.153	0.148	0.148
10	0.125	0.139	0.140	0.136	0.136
11	0.117	0.129	0.130	0.124	0.125
12	0.110	0.120	0.121	0.117	0.116
13	0.102	0.113	0.114	0.110	0.109
14	0.098	0.107	0.107	0.103	0.103
15	0.093	0.101	0.102	0.099	0.098

True: actual (empirical) probabilities, $\Pr\left(H_A \mid p_{(u)}\right) = 1 - \Pr\left(H_0 \mid p_{(u)}\right)$.
Estimated: probabilities computed by Equation (6) with prior $\Pr(H_0) = 1 - m/k$.
Estimated*: probabilities computed by Equation (6) with a corrected prior (Equation 17)

TABLE 2
Probability that at least one of five independent associations will rank among top-i false positives in a 700K GWAS

	$\Pr\left(\min(Y) \leq X_{(i)}\right)$			
Rank (i)	True	L_e-based	Median-based	Mean-based
1	0.193	0.195	0.158	0.183
10	0.413	0.415	0.409	0.414
25	0.538	0.537	0.535	0.537
50	0.635	0.635	0.634	0.635
100	0.726	0.730	0.729	0.730
200	0.815	0.815	0.816	0.815
500	0.904	0.905	0.904	0.904
1000	0.949	0.950	0.950	0.950
2000	0.977	0.978	0.978	0.978
5000	0.995	0.994	0.994	0.994

True: empirical ranking probabilities; $\boldsymbol{L_e}$**-based**: probabilities computed by Equation (14) using the MLE of L_e; **Median-based**: probabilities assuming independence, computed by Equation (15); **Mean-based**: probabilities assuming independence, computed by Equation (16).

TABLE 3
Probability that at least one association in the μ-opioid gene will rank among top-i false positives in a 700K GWAS

	$\Pr\left(\min(Y) \leq X_{(i)}\right)$			
Rank (i)	True	L_e-based	Median-based	Mean-based
1	0.191	0.194	0.168	0.187
10	0.345	0.346	0.343	0.345
25	0.426	0.426	0.425	0.426
50	0.494	0.492	0.491	0.492
100	0.561	0.562	0.561	0.562
200	0.631	0.632	0.631	0.631
500	0.725	0.723	0.723	0.723
1000	0.787	0.787	0.787	0.787
2000	0.846	0.844	0.844	0.844
5000	0.906	0.907	0.907	0.907

True: empirical ranking probabilities; $\boldsymbol{L_e}$**-based**: probabilities computed by Equation (14) using the MLE of L_e; **Median-based**: probabilities assuming independence, computed by Equation (15); **Mean-based**: probabilities assuming independence, computed by Equation (16).

TABLE 4

Probability that one true association will rank among top-i false positives in a two-million simulated SNP scan with extreme LD

	$\Pr\left(Y \leq X_{(i)}\right)$			
Rank (i)	True	L_e-based	Median-based	Mean-based
1	0.382	0.398	0.210	0.231
5	0.415	0.424	0.333	0.338
15	0.460	0.468	0.424	0.426
25	0.489	0.497	0.468	0.469
35	0.511	0.520	0.498	0.499
50	0.537	0.546	0.530	0.531
75	0.571	0.578	0.567	0.567
100	0.596	0.601	0.593	0.593
500	0.735	0.737	0.735	0.735
1000	0.791	0.792	0.791	0.791
5000	0.897	0.898	0.898	0.898

True: empirical ranking probabilities; $\boldsymbol{L_e}$**-based**: probabilities computed by Equation (14) using the MLE of L_e; **Median-based**: probabilities assuming independence, computed by Equation (15); **Mean-based**: probabilities assuming independence, computed by Equation (16).

FIGURES

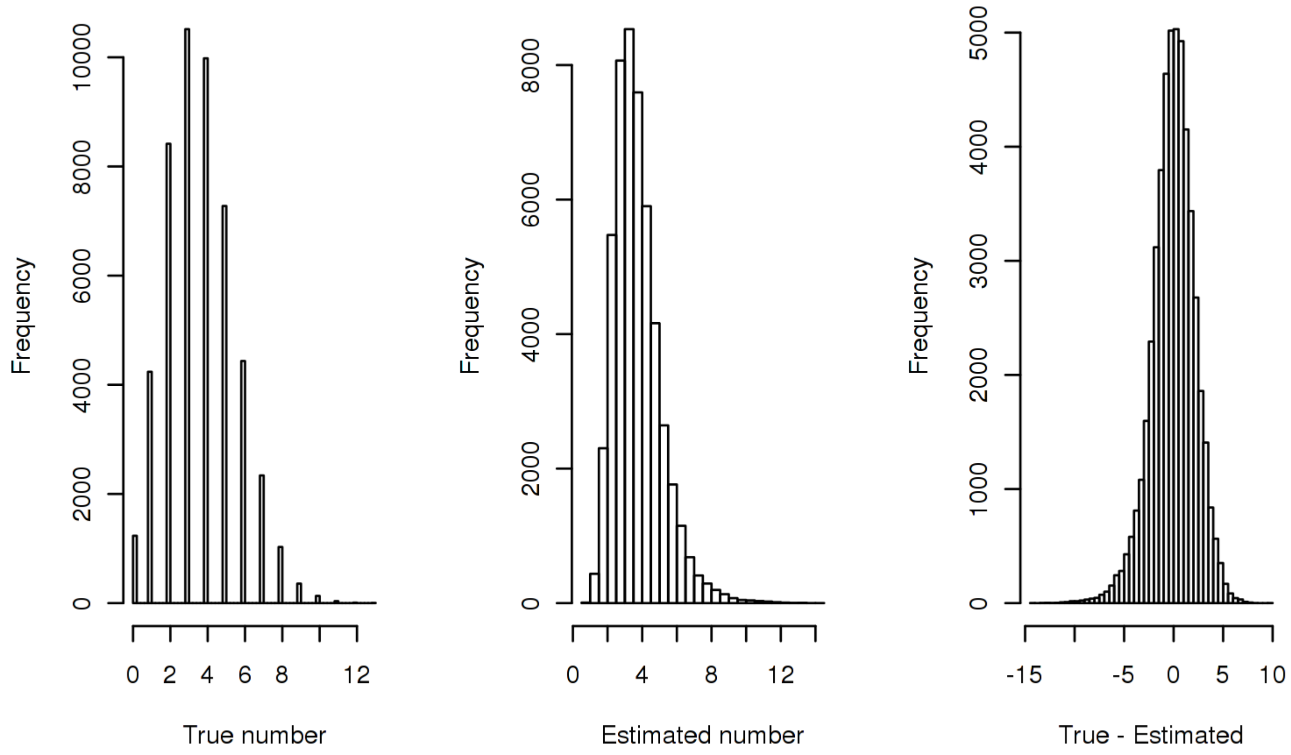

FIGURE 1 – Number of true associations among 15 best results in a two-million GWAS.

Figure legend: Left graph: the actual distribution. Middle graph: distribution estimated with Equation (6). Right graph: distribution of differences between the actual and the estimated number of true positives.