Title: Ranks of genuine associations in whole genome scans

(GENETICS; October 2005; in press)

Authors: Dmitri V. Zaykin[†] (zaykind@niehs.nih.gov), Lev A. Zhivotovsky[♯]

[†] National Institute of Environmental Health Sciences, National Institutes of Health.

[♯] N.I. Vavilov Institute of General Genetics, Russian Academy of Sciences, 3 Gubkin St., 117809, Moscow, Russia.

Running head:

Ranks of true positives in genome scans

Keywords:

Multiple testing, association tests, linkage disequilibrium, disease mapping

Corresponding author:

Dmitri Zaykin

National Institute of Environmental Health Sciences

MD A3-03, South Bldg (101), POB 12233

Research Triangle Park, NC 27709

Phone: (919) 541–0096; Fax: (919) 541–4311

email: zaykind@niehs.nih.gov

Abstract

With the recent advances in high-throughput genotyping techniques, it is now possible to perform whole-genome association studies to fine-map causal polymorphisms underlying important traits that influence susceptibility to human diseases and efficacy of the drugs. Once a genome scan is completed the results can be sorted by the association statistic value. What is the probability that true positives will be encountered among the first most associated markers? When a particular polymorphism is found associated with the trait, there is a chance that it represents either a "true" or a "false" association (TA vs. FA). Setting appropriate significance thresholds have been considered to provide assurance of sufficient odds that the associations found significant are genuine. However, the problem with genome scans involving thousands of markers is that the statistic values of FA's can reach quite extreme magnitudes. In such situations, the distribution corresponding to TA's and the most extreme FA's become comparable and significance thresholds tend to penalize TA's and FA's in a similar fashion. When sorting between true and false associations, it becomes important what is the "typical" place (i.e. rank) of TA's among the most significant outcomes, ordered by the association statistic value. The distribution of ranks that we study here allows calculation of several useful quantities. In particular, it gives the number of most significant markers needed for a follow-up study to guarantee that a true association is included with certain probability. This can be calculated conditionally on having applied a multiple testing correction. Effects of multilocus (e.g. haplotype association) tests and impact of linkage

disequilibrium on the distribution of ranks associated with TA's is evaluated and can be taken into account.

Introduction

Continuous efforts to characterize genetic contributions to human diseases led to identification of many susceptibility genes. Mapping of genes with well pronounced effects, such as involved in certain types of breast cancer, Alzheimer's disease and cystic fibrosis are examples of remarkable success. However, mapping complex diseases is generally complicated and many studies fail to replicate. Risch and Merikangas (1996) gave recommendations for sample sizes required to detect genetic effects of a specified size with certain power, $1-\beta$. If an association genome scan contains L markers, the corrected 5% type-I error rate is $\alpha = 0.05/L$, and the required sample size can be calculated for a normally distributed test statistic as $N = (Z_\alpha - \sigma Z_{1-\beta})^2/(2\mu^2)$, where μ, σ are the parameter values under the hypothesis of association, H_A and $Z_\alpha, Z_{1-\beta}$ refer to α and $1-\beta$ quantiles of the normal distribution (Risch and Merikangas, 1996). When high power at the genome-wide α-level can be assured, that will provide good odds that the most significant results represent actual associations.

An important question that has not been adequately addressed so far is the ranks of true associations. Distributions of ranks are particularly important in designing the follow-up studies and deciding how many markers are necessary to consider in a follow-up study in order to capture all true associations with a high probability (Satagopan *et al.*, 2002). Consider the following study as an example. Ozaki *et al.* (2002) established association of an SNP in the lymphotoxin-α gene with susceptibility to myocardial infarction. In a genome scan of 65,671 SNPs the actual association was originally found to have a p-value of

0.0022, being less significant than over 200 spurious associations that later failed to replicate. These included six false effects with p-values below 10^{-5} and one false effect with $p < 10^{-6}$. Is it a typical outcome – or would we rather expect genuine associations to be the front runners? Would TA's be usually ranking first among results that passed a "multiple testing correction"? The genome scan of Ozaki *et al* (2002) included 94 cases and 658 population controls. Their larger follow-up study determined the relative risk of about 1.6 associated with one of the genotypes at the putative associated SNP. An approach that we describe here allows us to estimate that there was only a 14% chance that the association with such an effect would rank among the first 200 most significant results. To cover such an effect with only 50% probability in a genome scan with 65,671 SNPs, one would need to examine over 3400 of the most significant results – a surprisingly high number. Even higher numbers are needed to cover the TA with certainty that is higher than 50%. Here we describe an approach that allows such calculations.

A simplified motivating description of the problem studied in this article is as follows. Suppose there is a single TA with a specific effect size, which could be measured by the penetrance value of a susceptibility allele. Given a fixed sample size, there is a certain probability (P) that the association test statistic calculated at this marker will exceed a given value, Z. However, as the total number of markers (L) increases, the largest test statistic among L false associations will eventually have a higher chance than P to exceed Z. In other words, the distribution associated with the the smallest "false" p-value will eventually become more skewed toward zero as L increases. In such situations, multiple testing corrections cannot distinguish among true and false associations and justly penalize the false ones. As a result, the position of TA's among results sorted by a measure of association or significance is not substantially affected (in a subset of experiments that passed the correction). Thus, outcomes of an association genome scan can be thought of as realizations of a stochastic experiment in which statistic values representing TA's and FA's compete in the magnitude of the value they are able to attain. Whichever statistic values ranked first may be associated with the highest odds of representing the TA's, and corresponding markers may become candidates for an independent follow-up study. Considering recent recommendations calling for relatively large (100K–200K) numbers of markers (Goldstein *et al.*, 2003) this becomes a problem of extensive multiple testing.

Traditional way to address the multiple testing issue is to establish significance thresholds that preserve probability of making a certain proportion of claims under the situation of ”complete null” (i.e. when all markers represent FA's). Lander and Kruglyak (1995) came up with explicit recommendations for significance thresholds that are appropriate for linkage studies where the correlation between adjacent markers is high and extends throughout long genetic distances. Some issues with the type-I error control approach are discussed below and its effect on the distribution of ranks is investigated. Multiple testing corrections to assure type-I error control require a statistic value (or equivalently a p-value) to pass a certain threshold to be declared ”significant”. The effect of this is merely a ”truncation” where an entire experiment (i.e. genome scan) could arguably be regarded a failure if none of the markers exhibit the threshold significance. We study the effect of such truncation on the relative order of ranks of true and false associations by restricting attention to the results among the experiments that pass the significance threshold. In a single given experiment, corrections such as Bonferroni do not change the order of results and the original ranks. However, the distributional characteristics of ranks will change for a random experiment given that some of the tests in that experiment have passed the correction. In other words, the distribution of the ranks is generally different for a subset of genome scans with some significant results, compared to all genome scans with and without significances.

A brief summary of our findings is as follows.

1. Investigation of relative TA and FA ranks allows to treat the multiplicity problem in a useful way that is complementary to more traditional type-I error and false discovery rate approaches. We find that in large-scale association genome scans, TA's are not likely to result in the largest corresponding association statistic values. This is in part due to small effect sizes associated with complex traits and corroborates concerns about the linkage disequilibrium (LD) mapping outlined in Terwilliger and Weiss (1998).

2. Multiple testing corrections do not have a drastic impact on the distribution of ranks, unless the power is high. Thus, procedures that preserve type-I error rate or the unconditional false discovery rate (FDR) are insufficient to guarantee that the most significant results represent TA's. Our approach helps to clarify sources of the apparently low chances of replication for originally found associations (Terwilliger and Weiss 1998; Vieland, 2001; Ioannidis *et al.*, 2001; Lohmueller *et al.*, 2003; Lucentini, 2004).

3. The effect of even strong LD on TA ranks is small to be of substantial importance in genome-wide association studies. This finding is in contrast with the statistical behavior in linkage analysis where the correlation extends throughout large chromosomal regions, greatly reducing the effective number of tests.

4. Dependency among the association test statistics due to block structure of LD has a similar small impact.

5. Multilocus approaches such as haplotype-trait association methods confer additional advantage in terms of the ranks of TA's. This finding is independent of previous arguments proposed in favor of haplotype analysis.

Ranks of true and false positives

It is possible to give explicit recommendations for genome-wide significance thresholds that provide appropriate type-I error control. However, type-I error control only considers the situation of the complete absence of actual associations. On the other hand, the question of practical importance is what are the chances that a particular test is a true association. Morton (1998) pointed out that it is desirable to assure high reliability (ρ) – the proportion of true discoveries among all, "true" and "false" discoveries, an essentially Bayesian concept. Reliability relates to the frequentist version of "false discovery rates" (FDR) of Benjamini and Hochberg (1995) as $\text{FDR} = E(1-\rho)\Pr(\mathcal{T}+\mathcal{F}>0) = E\left\{\mathcal{F}/(\mathcal{T}+\mathcal{F})\right\}\Pr(\mathcal{T}+\mathcal{F}>0)$, where $\mathcal{T}$ and $\mathcal{F}$ are the numbers of true and false associations that were declared significant. FDR is the average proportion of false positives across multiple studies, including those with no discoveries, and the reliability is approximating the expectation, $E\left\{\mathcal{T}/(\mathcal{T}+\mathcal{F})\right\}$, among the studies where one or more rejections have been made. The value $\Pr(\mathcal{T}+\mathcal{F}>0)$ decreases with the overall number of tests conducted within a study, and increases with power to detect TA's. The power is typically highly variable. Genome scan marker sets consist of common SNPs spread out throughout the genome. Actual causal polymorphisms are unlikely to be captured by any of the actual markers in the set. Instead, the hope is that the sufficient marker density may provide high LD with the causal loci. In this case some of the markers surrounding the functional polymorphisms can be viewed as proxies for TA's. However, the population-specific nature of LD, its high variability as well as the uncertainty in allele frequencies make it difficult to specify the magnitude of the association with a marker. Although high power at the genome-wide α-level will provide good odds that the most significant results represent actual associations, the actual power to detect associations in the vicinity of causal variants is likely to be low and indeterminate in practice. Expected *conditional* FDR, or the proportion of false discoveries among all discoveries, $E(1-\rho)$, can be related to the posterior probability of the null hypothesis (Storey, 2002). This approach is most straightforward in situations where the proportion of TA's is relatively high so that the distribution of mixture of TA's and FA's can be characterized empirically and distinguished from the distribution of FA's. More generally, it involves specifying the prior TA probability (expected proportion of TA's among all studied markers) as well as the power characteristics of markers representing the TA's. We study the problem in terms of the stochastic ordering and ranks of statistic values associated with true and false positives. Instead of calculating the probability of the null hypothesis (H_0) given a p-value, we compute the number of most significant results needed to contain a TA with a given confidence.

In the case of independence between genetic markers, the rank probabilities can be derived analytically. Next section describes necessary theory and Monte-Carlo approach used to approximate the analytic results as well as to model ranks under more complex situations of dependency due to LD.

Theory on stochastic ordering of true and false positives

To model the distribution of ranks, we start with the joint distribution of p-values corresponding to true and false associations, working in terms of order statistics that come from distinct densities corresponding to TA and FA p-values. In this regard, p-values are used as a ranking measure reflecting the degree of association. Equivalently, the problem can be described in terms of an arbitrary association measure used to order results. Integrating the joint distribution allows calculating quantities such as the probability that the largest true association p-value is smaller than the i-th smallest false association p-value, as well as the expected ranks for the true associations. These quantities can be calculated either unconditionally or conditionally upon a multiple testing correction and require the assumption of independence. A diffusion process is used to extend the method allowing for dependence due to LD between genetic markers. Consider a continuously distributed test statistic for association, T_i, at the marker i, with the corresponding p-value, $p_i = 1 - F_0(T_i)$, where $F_0(\cdot)$ is the cumulative distribution function (CDF) of T_i, assuming the null hypothesis, H_0. This is the correct CDF for markers representing FA's, but the actual CDF of the test statistic for TA's is denoted by $F_T(\cdot \mid \gamma)$. In the case of a χ^2-distribution, γ is the non-centrality parameter. If T_i are normally distributed, γ refers to the shift in the mean of the distribution caused by the association with the trait. Markers are ordered in terms of p-values, so that $\{p_1 \leq p_2 \leq ... \leq p_L\}$, where L is the overall number of markers in the study. For the FA's the CDF of p-values is $\mathcal{P}_0(p) \equiv \Pr(P \leq p) = p$, where P, p denote the random variable and its value respectively. For the TA's we have, assuming a one-tailed test

$$\mathcal{P}_T(p \mid \gamma) \equiv \Pr(P \leq p) = 1 - F_T(F_0^{-1}(1-p) \mid \gamma) \tag{1}$$

For illustrative purposes, we will now assume a normally distributed T_i, although calculations are similar for other continuous distributions. The results will also be presented assuming a χ^2-distribution which is more relevant in exploratory analysis so that the direction of the effect is not considered. With the normal T_i, equation (1) becomes $\mathcal{P}_T(p \mid \gamma) = 1 - \Phi\left(\Phi^{-1}(1-p) - \gamma\right)$, where $\Phi(\cdot)$ is the normal CDF, and $\gamma = \sqrt{N}\mu/\sigma$ is the power parameter that depends on the sample size, N, effect size, μ, and the variance, σ^2. The normal test statistic p-value density is $\pi_T(p \mid \gamma) = \frac{\partial \mathcal{P}_T(p|\gamma)}{\partial p} = \frac{\phi\left(\Phi^{-1}(1-p)-\gamma\right)}{\phi(\Phi^{-1}(1-p))}$, where $\phi(\cdot)$ is the normal density function. Note that $\mathcal{P}_0(p) = \mathcal{P}_T(p \mid \gamma = 0) = p$, and $\pi_T(p \mid \gamma = 0) = 1$ – i.e. $p \sim \text{uniform}(0,1)$ under H_0. Let X_i be the random variable corresponding to the i-th smallest FA p-value. Denote the most significant TA among m true associations, with its random p-value by $Y \equiv Y_1$. Consider events: $A \equiv Y < X_i$; $B \equiv \min(X_1, Y) \leq \delta$; $B^C \equiv \min(X_1, Y) > \delta$; where $\Pr(B) + \Pr(B^C) = 1$, and δ is a multiple-testing adjusted α-level, e.g. Šidak's adjusted δ is $1-(1-\alpha)^{1/L}$. For large L this gives $\approx \alpha/L$, the Bonferroni correction. If true and false effects are independent, the joint density function of X_i and Y is (dropping conditioning on γ and other parameters from the notation for simplicity)

$$\begin{aligned} &\pi_{X_i,Y}(x,y) = \pi_{X_i}(x)\ \pi_Y(y) \\ = \quad &\left\{\frac{x^{i-1}(1-x)^{s-i}}{B(i, s-i+1)}\right\} \left\{\frac{\phi\left(\Phi^{-1}(1-y)-\gamma\right)}{\phi\left(\Phi^{-1}(1-y)\right)} m \left[\Phi\left(\Phi^{-1}(1-y)-\gamma\right)\right]^{m-1}\right\} \end{aligned} \tag{2}$$

where the number of FA's is $s = L - m$, $B(\cdot)$ is the beta function, the first ratio is the beta$(i, s-i)$ distribution of the i-th smallest false p-value, and the ratio of normal densities is the density for the TA p-value distribution, assuming independence and common power characteristics for each of m associations. The problem is to find $\Pr(A \mid B) = \Pr(Y \leq X_i \mid (Y \text{ or } X_1) \leq \delta)$. First consider $m = 1$. $\Pr(A \mid B)$ is the probability that the true association ranks below the i-th most significant FA p-value. It is found as

$$\begin{aligned} \Pr(Y \leq X_i \mid (Y \text{ or } X_1) \leq \delta) &= \frac{\Pr(A) - \Pr(A \cap B^C)}{1 - \Pr(B^C)} \\ = & \frac{\int_0^1 \int_y^1 \pi_{X_i,Y}(x,y)\ dx\, dy - \int_\delta^1 \int_y^1 \int_\delta^x \pi_{X_1,X_i,Y}(z,x,y)\ dz\, dx\, dy}{1 - \int_\delta^1 \int_\delta^1 \pi_{X_1,Y}(x,y)\ dx\, dy} \end{aligned} \tag{3}$$

where

$$\pi_{X_1,X_i,Y}(z,x,y) \quad = \quad \left\{ \frac{s!\ (x-z)^{i-2}(1-x)^{s-i}}{(i-2)!(s-i)!} \right\} \left\{ \frac{\phi\left(\Phi^{-1}(1-y) - \gamma\right)}{\phi\left(\Phi^{-1}(1-y)\right)} \right\} \tag{4}$$

The first term in (4) is the joint distribution of X_1 and X_i, independent of Y with density given by the second term. Alternatively, we would like to find a value of i such that $\Pr(A \mid B)$ is equal to a certain value, such as 95%. This is found by computing (3) over the range of i. The resulting value of i is equivalent to the number of most significant associations that will contain the actual TA with 95% probability. Equation (3) is measuring the probability that the true positive will rank right below the i-th most extreme false positive. This probability is conditional on at least one p-value satisfying the multiple testing threshold, δ. An unconditional version is obtained by setting $\delta = 1$. Satagopan *et al.* (2002) considered the unconditional ranking in the problem of optimizing the total number of typed markers for two-stage association mapping study designs. The "average" ranking of Y, or its expected rank is given by

$$E\left\{\Pr(A \mid B)\right\} \;=\; L - \sum_{i=1}^{L} \Pr(Y \leq X_i \mid (Y \text{ or } X_1) \leq \delta) \tag{5}$$

To modify equation (3) allowing for the calculation of probability that *all* TA's will rank below i, we need the distribution of the largest TA p-value, among $(m > 1)$ TA's:

$$\pi_{Y_m}(y) = \frac{\phi\left(\Phi^{-1}(1-y) - \gamma\right)}{\phi\left(\Phi^{-1}(1-y)\right)} m \left[1 - \Phi\left(\Phi^{-1}(1-y) - \gamma\right)\right]^{m-1}$$

as well as the joint distribution of the maximum and minimum p-value for TA's:

$$\begin{aligned} \pi_{Y_1,Y_m}(u,v) &= m^3(m-1)\left[1 - \mathcal{P}_T(u)\right]^{m-1} \mathcal{P}_T(v)^{m-1} \\ &\times \left(\left[1 - \mathcal{P}_T(u)\right]^m + \mathcal{P}_T(v)^m - 1\right)^{m-2} \pi_T(u)\pi_T(v) \end{aligned}$$

Probability that all true positives will rank before the first i false positives, conditional on a multiple testing correction is:

$$\begin{aligned} & \Pr(Y_m \leq X_i \mid (Y_1 \text{ or } X_1) \leq \delta) \\ = & \frac{\int_0^1 \int_y^1 \pi_{X_i}(x)\pi_{Y_m}(y)\ dx\, dy}{1 - \int_\delta^1 \int_\delta^1 \pi_{X_1}(x)\pi_{Y_1}(y)\ dx\, dy} \\ - & \frac{\int_\delta^1 \int_{y_m}^1 \int_\delta^{y_m} \int_\delta^{x_i} \pi_{X_1,X_i}(x_1,x_i)\pi_{Y_1,Y_m}(y_1,y_m)\ dx_1\, dy_1\, dx_i\, dy_m}{1 - \int_\delta^1 \int_\delta^1 \pi_{X_1}(x)\pi_{Y_1}(y)\ dx\, dy} \end{aligned} \tag{6}$$

The expected rank of Y_m is obtained in the same manner as in (5).

These equations can be evaluated by numerical integration or more conveniently via Monte-Carlo simulations. Simulations are set up as follows. FA's are sampled from the uniform(0,1) distribution. The TA p-value distribution function, $\mathcal{P}_T(\cdot)$, can be inverted to sample each of m true association p-values. For example, assuming a normally distributed test statistic, $\{p_i\}$ are obtained as $p_i = 1-\Phi\left(\Phi^{-1}(u_i)+\gamma\right)$, where u_i is a uniform(0,1) random deviate. With a chi-square test statistic, p-values are generated as $p_i = 1 - \Psi_{d,0}\left(\Psi^{-1}_{d,\gamma}(u_i)\right)$, where $\Psi_{d,\gamma}(\cdot)$ is the χ^2 CDF with d degrees of freedom and the non-centrality parameter γ. Once all $L = s + m$ of p-values are generated, they are ordered and their ranks are recorded. This consists of a single simulation experiment. Probabilities such as in (3) are approximated by the proportion of times across multiple experiments where Y had ranked below the i-th false positive. Parameter γ governs the power associated with individual TA's. For example, γ=3.29 with the normally distributed test statistic corresponds to the 95% probability of detecting true effects with a 5%-level test, and can be obtained as $\Phi^{-1}(1-0.05)+\Phi^{-1}(0.95) \approx 3.29$. Such simulations are easy to set up and they become more convenient than the numerical integration when L is large. Although we considered a fixed value of m for the calculations to be more transparent, in principle this value can be allowed to have a distribution allowing for uncertainty associated with m.

In this study, we applied the Monte-Carlo approach to obtain numerical results taking the number of simulations to be at least 50,000. Simulation approach is especially useful to study the effect of correlations between p-values. The positive dependence, such as dependence due to LD is expected to reduce the effect of the total number of false effects on the distribution of ranks of TA's. Assuming a multivariate normal distribution for the test statistic and the exponential correlation decay with distance, the joint distribution of the test statistic under H_0 over neighboring markers can be described by the Ornstein-Uhlenbeck diffusion process, $\frac{dZ}{dt} = -aZ(t) + \sigma\tilde{\xi}(t)$, where $\tilde{\xi}(t)$ is the white noise term and a, σ^2 are the process drift and the variance parameters. After statistic values (normal scores) are converted to p-values, sampling from this diffusion process generates correlated p-values with the stationary uniform (0,1) distribution. The correlation $r(Z_i, Z_j)$ generated by diffusion closely translates to the correlation between p-values: $r(p_i, p_j) = 6 \ \arcsin\left[r(Z_i, Z_j)/2\right]/\pi$ (Kruskal, 1958). The largest ratio of $r(Z_i, Z_j)$ to the correlation between two p-values is $\pi/3 \approx 1.047$, as $r(Z_i, Z_j)$ approaches zero. We have also implemented a more realistic model of the correlation structure allowing for extended blocks of very high LD interspersed with regions of low LD, by mixing two (high and low correlation) Ornstein-Uhlenbeck diffusions. In reality, p-value correlations may not be adequately described by the diffusion. Nevertheless, this model allows to generate substantial correlations with specified decay characteristics and therefore allow investigation of the effect of correlation on the distribution of TA ranks.

Results

We considered two basic calculations. The first is obtaining the probability that TA's are found among some fixed number (R) of most significant results in a scan (Table 1). The second calculation is the number of most significant results required to contain TA's with some fixed probability (Tables 2–6). Both calculations are performed with or without a multiple testing correction. Imposing a multiple testing correction results in discarding experiments that did not pass the significance threshold. Then the ranks are calculated for the subset of all scans that have passed the correction. Although the ranks do not change for a single given genome scan, they are expected to improve among the scans that passed the threshold. This is modeling the expectation of a researcher that genome scans with multiplicity-corrected, statistically significant results are more likely to represent TA's, especially among the tests that passed the correction. The effect of LD on the ranks is studied for the case of strong LD that is homogeneous across the genome, as well as for the case of LD that has a block structure. For the case of multiple TA's, chances of finding some or all of the effects are evaluated. Finally, the effect of multilocus tests (e.g. haplotype association tests) is considered and compared to the results obtained for tests based on individual SNPs.

Table 1 shows unconditional and conditional probabilities of finding all $m = 3$ TA's among R most significant results in a genome scan with s=100,000 independent FA's. We used four different values of power to detect individual true effects (0.75, 0.85, 0.95, 0.99) assuming the normal distribution in (1), and compared probabilities under no multiple testing correction

($\delta = 1$) with probabilities obtained conditional on at least one p-value satisfying the Šidak correction, $\delta = 1-(1-0.05)^{1/100003}$. Table 1 indicates that with a multiple testing correction, the chances that all three true effects are found among R smallest p-values are improved by only a small amount. With 95% power, we need to consider around 500 most significant results to bring this probability up to 44%. Multiple testing correction only slightly increases it to 54%. This improvement would be even less pronounced for adjustments that are less stringent than Šidak or Bonferroni. Power has the most definite influence on the probability to find true effects. When the power to find any one of the three true effects in a single 5%-level test is 75%, there is practically no chance of finding all three effects when up to 100 markers are considered. The probability conditional on the multiple testing adjustment is still small even with R=1000 for this value of power. The effect of the power associated with TA's is of great importance. There is considerable improvement in probability to find true effects when power reaches values of 99% and above. For example, the increase in power from 95% to 99% lowers R from 500 to 50 markers needed for 50% probability of discovery. This effect is also apparent in Tables 2 and 3. These tables use s=200,000 of FA's and a single TA ($m = 1$) assuming the $\chi^2_{(1)}$ distribution for the test statistic in (1). The probability that this effect is among first R most significant markers is set to 0.95 for Table 2 and to 0.50 for Table 3. The entries in the tables show the value of R required to satisfy these probabilities. Dependency among p-values due to LD is added, assuming 200,000 equally spaced markers with correlation decay characteristics as shown in Figure 1. The correlation shown in Figure 1 corresponds to a quite extensive LD. Generally, high correlation values between alleles at two loci (i.e. LD) translate into much weaker correlation between the corresponding p-values computed for two single-locus tests of association (Nielsen *et al.*, 2004). Tables 2 and 3 show that unless the power is very high, a substantial proportion of the original 200,000 marker set needs to be examined to ensure that it contains the TA. The effect of power is most substantial at values close to 99%. Only at these high power values there is a definite benefit of a multiple testing correction, in that the rank of the true association lowers to manageable numbers (e.g. 1671 markers without the correction vs. 70 markers for 99% power and no-LD entry in Table 2). Since LD usually follows a block structure, additional simulations were carried out with the mixture of two diffusions. The decay of correlation in blocks and between the blocks followed the pattern showed in Figure 2, with 25% of SNPs allocated within the blocks. Proportions of genome regions within blocks have been reviewed by Wall and Pritchard (2003) for various populations and subsets of genome. The usage of p-value correlation most closely corresponds to the LD measured by the composite correlation r_{AB} (Weir, 1996) which translates into a stronger requirement than that of D' - the gametic or the composite LD normalized by the bounds on covariance between alleles (Lewontin, 1964, Hamilton and Cole, 2004, Zaykin, 2004). The coefficient r_{AB} has well-defined statistical and population-genetic properties. It serves well in contexts of association mapping (Meng *et al.*, 2003), because high values of r_{AB} imply that an allele at one locus can be regarded as a proxy for an allele at another locus. This requires dependency as well as the relative closeness of allele frequencies at both loci.

The block LD parameters were chosen to result in the correlation decay similar to the one shown in Figure 1, when averaged across large distances. Results shown in Table 4 indicate that the block structure of correlation did not substantially affect the outcomes compared to the results obtained for the monotone decay of correlation with distance. The effect of dependency among markers has a very small effect on the ranking of TA's, because of the local nature of the correlation decay. As far as ranks of true positives are concerned, the "effective number of tests" is roughly equal to the total number of markers assuming no LD. It is our experience that in actual genome-wide association studies the average correlations between association test p-values are smaller in magnitude, with the decay that resembles the shape used in the present study.

When the number of TA's is increased, the probability to find *all* TA rapidly decreases. For example, with $m = 1$ and 99% power, the number of most significant markers that contain the TA with 95% probability is 1671 (with no correction) and 70 (with Šidak correction, Table 1, independence case). When $m = 2$, the numbers increase to 3969 and 2155, respectively. With $m = 3$, they become 6112 and 4573. These calculations assume the same 99% power for all TA's, although it is likely that as m increases, the power associated with individual TA's decreases.

The probability of finding *one or more* of TA's increases with the number of TA's but only under the unlikely assumption that the power associated with individual effects does not decrease. With $m = 1$ and 78.5% power, the number of most significant markers that contain the TA with 95% probability is 54233 (with no correction) and 47620 (with Šidak correction). With $m = 3$, the numbers become 3165 and 2196, although it could be naively assumed that these should correspond to the case with $m = 1$ and $1 - (1 - 0.785)^3 = 99\%$ power. More results for the case of $m = 3$ are given in Table 5. The results confirm that lower power associated with each effect is sufficient to obtain the same probabilities as those with a single false effect. For example, 80% power associated with each of 3 TA's results in similar numbers as that for a single TA with 95% power (for 50% probability of containing TA's). It is expected however that given multiple effects the power associated with each of them will generally be quite low. An assessment of the anticipated decrease in power can be obtained under a simple model when multiple effects independently contribute to a binary phenotype. Suppose there are m susceptibility polymorphisms contributing to the trait. If $m = 2$ with two corresponding frequencies among the cases, p_1, p_2, then the frequency of either one of the susceptibility polymorphisms is $p_c = p_1 + p_2(1 - p_1) = 1 - (1 - p_1)(1 - p_2)$. For simplicity, assume the same common frequency of all polymorphisms among the cases, p. Then $p_c = 1-(1-p)^m$, and conversely $p = 1-(1-p_c)^{1/m}$. Similar calculation would hold for the frequency among controls, q_c. For a given relative risk p_c/q_c it is possible to calculate the expected power associated with any one of these m polymorphisms. Resulting power lines using the relative risk power formula (equation 7, discussed below) are shown in Figure 3 for p_c=0.3 and q_c=0.2. The discrepancy is high - nearly 100% power of the compound test for all three effects has the power of about 85% for the individual effects. In other words it is difficult to assure that high power for all three effects (e.g. last row in Table 5) is satisfied in practice.

Genome scan analysis may involve multilocus tests, e.g. statistical tests relating haplotype frequencies with phenotype values. There is a problem of which polymorphisms are to be included in a particular test. Previously, we investigated power of the "sliding window" approach where several neighboring polymorphisms are included into the current window and the overall test is performed (Zaykin *et al.*, 2002a). The total number of tests for the whole genome scan remains essentially the same although there is additional but very short-distance correlation between the tests sharing the same polymorphisms. Such multilocus approaches are advantageous when there are substantial haplotype effects. Nevertheless, there is the problem of balance between the increase in degrees of freedom (d.f.) associated with the haplotype tests and the increase in the effect size – no power increase is guaranteed with the haplotype analysis even if the effects are specifically haplotype-driven. Another possible advantage of the haplotype analysis is that haplotypes can be in higher LD with unobserved mutations than the individual SNPs comprising the haplotypes. However, in such situations haplotype tests rarely result in substantial gain in power. Higher power compared with tests for individual SNPs can be observed primarily when there are multiple susceptibility SNPs and the LD is low (Morris and Kaplan, 2002). In our framework, the major influence of the multilocus (e.g. haplotype) tests on the distribution of ranks is via the form of the distribution of p-values corresponding to multilocus TA's. We considered the situation of block-LD correlated test statistics with eight d.f. that would for example correspond to the overall genotypic test involving two diallelic SNPs, with nine distinguishable dilocus genotypes. A haplotypic test with three SNPs would have a similar number (seven) of d.f. Results are shown in Table 6. Although the results are similar to the one d.f. tests from Table 4, all eight d.f. ranks are smaller than those for the one d.f. tests (excluding the situations when the power is high enough that the TA ranks first). The difference between the one and eight d.f. tests is more pronounced under the multiple-testing correction (second and fourth columns of the tables). The difference between entries of two tables is statistically significant (Wilcoxon signed-rank test p-value is 0.0005 for comparing just 95% entries). The relation between the d.f. (assuming a chi-square distribution of the association measure) and the distribution of ranks is rather complicated. For a particular quantile of the p-value distribution the exact relation regarding the optimal d.f. can be based on the following computation. Let $(\gamma_1, \gamma_2, ..., \gamma_k)$ be the non-centrality parameters corresponding to $(1, 2, ..., k)$ d.f. tests to have the same power $(1 - \beta)$ at a particular α-level. Then $i = 1, ..., k$ significance levels α_i^* at the quantile q are $\alpha_i^* = 1 - \Psi_{d_i,0}\left(\Psi_{d_i,\gamma_i}^{-1}(1-q)\right)$, where $\Psi_{d_i,\gamma_i}(\cdot)$ is the χ^2 CDF with d_i degrees of freedom and the non-centrality parameter γ_i. A sample graph for 80% power tests and the 70% quantile is shown in Figure 4. Note that at 80% quantile the graph would be the straight line at 0.05. In this figure, the four d.f. test appears “optimal”, as it corresponds to the minimum significance level, however more important is the relation at smaller quantiles, somewhere in the neighborhood of the smallest expected p-values corresponding to false effects (i.e. around the Bonferroni level). At such level (with 200K tests) the “optimal” d.f. is equal to nine (Figure 5) and even much larger d.f. tests still have smaller significance levels than a single d.f. test. In general, the exact relation between the d.f. and the significance level depends on a particular quantile and the power at a given α-level. Still, our results indicate that tests with moderate degrees of freedom, e.g. haplotype tests including two to four SNPs may score better in terms of the ranks. This provides an additional justification for multilocus or haplotype-based analysis used in whole genome scans.

Returning to the single-SNP analysis of Ozaki *et al.* (2002), we emphasize the low (14%) estimated probability that their associated SNP with the estimated relative risk of 1.6 would rank among the first 200 most significant results, like have been observed in their study. For that SNP to appear with 50% probability among the most significant results, the number of SNPs is estimated to be 3460, given their sample size of 94 cases and 658 controls. For the case and the control frequencies (p, q), the power of a test for $H_0 : RR = 1$ can be obtained by considering the asymptotic normal distribution of $\ln(p/q)$. For example, equation (27) in Zaykin *et al.* (2004) can be inverted to obtain

$$\Pr(P < \alpha) = \Phi\left(\sqrt{N}\ln(p/q)/\sqrt{1/p + 1/q - 2} - \Phi^{-1}(1 - \alpha/2)\right) \tag{7}$$

This equation is for the equal sample size of the cases and the controls, but an "effective common sample size" can be calculated that would provide the same power as two given sample sizes of cases and controls. The power curve closely follows that of the more common test for the difference between two frequencies (equation 3 in McGinnis *et al.*, 2002). For two sample sizes, n_1 and n_2, we define the "effective common sample size" as the sample size N that would provide approximately the same power using such test. We find the value of N as

$$N = \frac{p(1-p) + q(1-q)}{\frac{p(1-p)}{n_1} + \frac{q(1-q)}{n_2}} \tag{8}$$

When $p = q$, the common sample size N in (8) is the harmonic mean of n_1 and n_2. With the relative risk of 1.6, the power reaches 99% when $n_1 = n_2 = 990$. With such sample size, there is 50% probability that the associated SNP would appear among first 2 most significant results, and 90% probability that it would appear among first 200. Reports of large-scale association studies are still scarce. One such recent study (Kammerer *et al.*, 2004) identified a replicated association on chromosome 19 in an intercellular adhesion molecule gene (ICAM) that influence nonfamilial breast cancer risk. Kammerer *et al.* used a pooled DNA genome-wide scan with 25,495 SNPs and a sample of 254 cases and 268 controls. Disregarding some increase in the test statistic variance due to pooling, we estimated that there was 84% probability to encounter the true positive with the relative risk of 1.3 reported in their study among first 550 results. We estimated the mean rank of the true positive being 506. The actual rank of that SNP in the genome scan of Kammerer *et al.* was 550, which is in close agreement with our prediction (Matt Nelson, personal communication).

Discussion

True associations are difficult to find even when they are present among the results, as they tend to rank quite high among all results ordered by the magnitude of a measure of association or by p-values. Common solution of stringent multiple testing control fails to influence the relative order of true and false associations, unless the power to detect true associations is very high. When viewed as a problem of stochastic ordering, the nature of this becomes clear. Multiple testing procedures penalize collections of p-values (whole genome scans) in a similar fashion, with no regard to which genome scans produced true associations ranking at the beginning of the results. Introducing high level of dependency only slightly reduces the effective number of markers, because the extent of the average correlation is short on the whole genome scale. It is essential that sufficient sample sizes are used to ensure very high power to detect true associations. This problem is specific to situations where the number of tests is very large and approximately equal to the number of false effects ($L \approx s \gg m$), such as in association genome scans. At the extreme of only two markers with one of them being a true association, the probability of correctly identifying it using a p-value quickly increases with power. On the other hand, whole genome scans with more than 100,000 markers demand extremely high power values (at least 99%) in order to reduce the number of markers required for replication in a follow-up study to a manageable value. Probability to locate a true association increases more rapidly for the high values of power. The increase from 95% to 99% is of much greater consequence than the increase from 85%

to 89%, although it is the the relative changes e.g. (1-0.95)/(1-0.99) vs. (1-0.85)/(1-0.89) that are more comparable.

Our examples assume much lower power values than that used in calculations of Risch and Merikangas (1996), who considered sample sizes needed to achieve the power of 80% at the α-level of $0.05/L$. The power values we use correspond to $\alpha = 0.05$, reflecting the influence of the effect magnitude, variability, and the sample size at a given marker. So, for example 99.999% power at $\alpha = 0.05$ corresponds to the power of 80% at $\alpha = 0.05/200000$ for a normally distributed test statistic. Such high power values will demand very large sample sizes, and high power at the genome-wide α-level will guarantee that TA's are likely to appear as first runners among the results. Practically, sample sizes continue to be modest, especially with regard the generally small sizes of genetic effects associated with complex traits. Moreover, genome scan SNPs are only expected to be proxies for actual causative polymorphisms through highly variable and population-specific LD that further decreases the effect size of such indirect associations. When multiple moderate effects contribute to the trait variation it is tempting to combine results from markers with large statistic values. However, our results suggest that the TA's tend to be interspersed with hundreds to thousands of false effects and the results of such analysis should be viewed with caution.

One could look at the issue of the number of tests (L) vs. power (that depends on $\sqrt{N}$) trying to come up with a way to design an optimal configuration of L and N. The problem is that the high marker density is required due to relatively short extent of LD in human populations, thus high values of L (100,000 or more) are needed (Goldstein *et al.*, 2003).

Therefore, if only a certain number of markers can be followed up in a replication study, the calculations for the required N should be carried out. If a case-control test is conducted, a quantity of interest can be the relative risk (RR) of a genetic variant, A. If the frequencies of A among the cases and controls are p, and q, respectively, the power of a test for $H_0 : RR = 1$ is approximately as given in (7). The population frequency of A is $wp + (1 - w)q$, where w is the prevalence of cases in the population. The same power calculation corresponds to a test that the genetic susceptibility associated with A is the same as the population prevalence (Zaykin *et al.*, 2004).

A potential issue with the TA definition is that the significance around false effects might be relatively more erratic, and there might be some information in the markers surrounding the peak corresponding to a TA. However, this problem remains controversial. In the context of linkage analysis Terwilliger *et al.* (1997) claimed that the peaks around true positives are expected to be wider. However Visscher and Haley's (2001) re-examination suggested that there is no additional information in the length of the peak. Siegmund (2001) concluded that in general methods that take the length of the peak into account (e.g. various smoothing techniques) can be somewhat helpful but are not expected to achieve substantial gains and that statistically this problem remains difficult. We have previously proposed an approach to combine neighboring p-values in a sliding window by the "Truncation Product Method" (Zaykin *et al.*, 2002b). The result of an application of this method is a new set of p-values combined (smoothed) over regions covered by the sliding window. Theoretically, a combined p can exceed the values of individual p-values in the combined set.

We found that multiple testing corrections have only mild effect on the ranks of the TA's. That is, among genome scans that yield one or more tests satisfying the significance threshold, the ranks of TA's (their expected distributional characteristics) are only moderately improved. This is in line with the observation that the proportion of false discoveries among multiple-testing corrected results can be substantially higher than the significance level (Zaykin *et al.*, 2000). High amount of LD between the markers or the presence of the LD block structure have not been found to greatly influence the ranks of TA's. The extent of high correlation has to be large enough to span whole segments of chromosomes for the effect to be considerable. In the context of linkage analysis such extended correlation has profound effect as has been demonstrated previously (Lander and Botstein, 1989).

Our results suggest that statistical tests with moderately large degrees of freedom (e.g. haplotype or other multilocus tests) may score better in terms of the ranks compared to single SNP tests with one or two degrees of of freedom. When either of the tests has the same power at a conventional significance level, the moderate degrees of freedom tests tend to rank closer to zero in terms of the p-values. This implies lower conditional FDR for these tests and gives an additional justification for using haplotypic tests in whole genome scans. The exact recommendations are difficult to come up with, as the results depend on parameters other than the degrees of freedom, as described above. In genome scans with over 100,000 markers, tests with 6 or higher d.f. are expected to have better or equal distribution of TA ranks than tests involving single SNPs.

In summary, we suggest that actual genetic associations are unlikely to appear among the front runners, unless high power at genome-wide level can be assured. We emphasize that although the results may be ”type-I error protected”, the most significant observations are still likely to represent false positives. Similar observations are being found in actual association scans (Ozaki *et al.*, 2002, Kammerer *et al.*, 2004). Considering substantial efforts put in every association genome scan as well as typically high prior understanding of genetic contribution to the variation in the trait of interest, it is important to consider the number of markers that should be subject of careful consideration in an independent follow-up study. We provide an approach to make such calculations. A program for computing true association ranking probabilities is available from DVZ upon request.

Acknowledgments

Clive Bowman, Meg Ehm, Ralph McGinnis, Mike Mosteller, Matt Nelson, Shyamal Peddada, Liling Warren and two anonymous reviewers provided useful comments that improved the manuscript.

References

Benjamini Y, Hochberg Y 1995 Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society series B-methodological, **57:** 289–300.

Goldstein DB, Ahmadi KR, Weale ME, Wood NW 2003 Genome scans and candidate gene approaches in the study of common diseases and variable drug responses, Trends in Genetics **19**: 615-622.

Hamilton DC, Cole DE 2004 Standardizing a composite measure of linkage disequilibrium. Ann Hum Genet. **68:** 234–239.

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. 2001 Replication validity of genetic association studies. Nat Genet. **29:** 306–309.

Kammerer S, Roth RB, Reneland R, Marnellos G, Hoyal CR, Markward NJ, Ebner F, Kiechle M, Schwarz-Boeger U, Griffiths LR, Ulbrich C, Chrobok K, Forster G, Praetorius GM, Meyer P, Rehbock J, Cantor CR, Nelson MR, Braun A. 2004 Large-scale association study identifies ICAM gene region as breast and prostate cancer susceptibility locus. Cancer Res. **64:** 8906–8910.

Kruskal WH 1958 Ordinal measures of association. J Amer Statist Assoc. **53:** 814–861.

Lander ES, Botstein D 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **21:** 185–199.

Lander ES, Kruglyak L 1995 Genetic dissection of complex traits: guidlines for interpreting and reporting linkage results. Nat Genet **11:** 241–247.

Lewontin RC 1964 The interaction of selection and linkage. I. general considerations; heterotic models. Genetics **49:** 49–67.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN 2003 Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet **33:** 177–182.

Lucentini J 2004 Gene association studies typically wrong. The Scientist **18:** 20.

McGinnis R, Shifman S, Darvasi A 2002 Power and efficiency of the TDT and case-control design for association scans. Behav Genet. **32:** 135–144.

Meng Z, Zaykin DV, Xu CF, Wagner M, Ehm MG 2004 Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. Am J Hum Genet **73:** 115-130.

Morris RW, Kaplan NL 2002 On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epid. **23**: 221–233.

Morton NE 1998 Significance levels in complex inheritance. Am J Hum Genet **62:** 690–697.

Nielsen DM, Ehm MG, Zaykin DV, Weir BS 2004 Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Genetics **168:** 1029–1040.

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T 2002 Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat Genet **32**: 650–654.

Risch N, Merikangas K 1996 The future of genetic studies of complex human diseases. Science **273:** 1516–1517.

Satagopan JM, Verbel DA, Venkatraman ES, Offit KE, Begg CB 2002 Two-stage designs for gene-disease association studies. Biometrics **58:** 163–170

Siegmund D 2001 Is peak height sufficient? Genet Epid **20:** 403–408

Storey JD 2002 A direct approach to false discovery rates. Journal of the Royal Statistical Society, Series B, **64:** 479–498.

Terwilliger JD, Shannon WD, Lathrop GM, Nolan JP, Goldin LR, Chase GA, Weeks DE 1997 True and false positive peaks in genomewide scans: Applications of length-biased sampling to linkage mapping. Am J Hum Genet. **61:** 430–438.

Terwilliger JD, Weiss KM 1998 Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol. **9:** 578–94.

Vieland VJ 2001 The replication requirement. Nat Genet. **29:** 244–245.

Visscher P, Haley C 2001 True and false positive peaks in genomewide scans: The long and the short of it Genet Epid. **20:** 409–414.

Wall JD, Pritchard JK 2003 Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet. 2003 **4:** 587–597.

Weir BS 1996 Genetic Data Analysis II. Sinauer Associates, Sunderland, MA.

Zaykin DV, Young SS, Westfall PH 2000 Using the false discovery rate approach in the genetic dissection of complex traits: a response to Weller *et al.* Genetics **54:** 1917–1918.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG 2002a Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered. **53:** 79–91.

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. 2002b Truncated product method for combining P-values. Genet Epidemiol. **22:** 170–185.

Zaykin DV 2004 Bounds and normalization of the composite linkage disequilibrium coefficient. Genet Epidemiol **27:** 252–257.

Zaykin DV, Meng Z, Ghosh SK 2004 Interval estimation of genetic susceptibility for retrospective case-control studies. BMC Genetics 2004, **5:** 9.

Conditionality	No correction: $\delta = 1$				Šidak: $\delta = 1 - (1 - 0.05)^{1/L}$			
Power	75%	85%	95%	99%	75%	85%	95%	99%
Value of R								
3	0.0000	0.0003	0.0044	0.0505	0.0000	0.0009	0.0121	0.0849
5	0.0000	0.0005	0.0125	0.1064	0.0015	0.0033	0.0326	0.1718
10	0.0006	0.0024	0.0306	0.1897	0.0019	0.0073	0.0698	0.2702
25	0.0020	0.0082	0.0714	0.3168	0.0039	0.0207	0.1336	0.4057
50	0.0043	0.0163	0.1202	0.4294	0.0087	0.0363	0.1984	0.5099
100	0.0101	0.0342	0.1888	0.5397	0.0186	0.0653	0.2825	0.6119
350	0.0414	0.1173	0.3719	0.7288	0.0606	0.1625	0.4770	0.7860
500	0.0612	0.1562	0.4402	0.7800	0.0780	0.2071	0.5374	0.8285
1000	0.1168	0.2630	0.5775	0.8613	0.1493	0.3201	0.6528	0.8911

Table 1: Probabilities that all $m = 3$ true effects are found among R smallest p-values (with 100,000 false effects)

Power	IID	IID	LD	LD
	No correction	Šidak	No correction	Šidak
60%	109311	108011	108941	107667
70%	81367	75476	79006	74593
80%	50021	43611	48637	41539
90%	21686	13924	20447	12230
95%	10002	3555	9931	2936
99%	1671	70	1604	49

Table 2: Number of most significant SNPs that will contain a single true association with 95% probability (1 TA + 200,000 FA markers scan). ”IID” – independent, identically distributed tests; ”LD” – assuming monotone LD between the tests with decay characteristics as in Figure 1.

Power	IID	IID	LD	LD
	No correction	Šidak	No correction	Šidak
60%	5298	4883	5165	4803
70%	2662	1992	2632	1742
80%	1005	500	1000	416
90%	242	15	240	9
95%	63	1	59	1
99%	5	1	5	1

Table 3: Number of most significant SNPs that will contain a single true association with 50% probability (1 TA + 200,000 FA markers scan). "IID" – independent, identically distributed false effects; "LD" – assuming monotone LD between the tests with decay characteristics as in Figure 1.

Chances to contain TA	50%		95%	
Power	No correction	Šidak	No correction	Šidak
60%	5363	4056	109502	105410
70%	2564	1221	78919	69761
80%	1037	203	50795	37167
90%	232	5	22216	11042
95%	59	1	9904	2311
99%	3	1	1620	40

Table 4: Block LD dependency with one true effect ($m = 1$). Entries represent the number of most significant SNPs that will contain a single true association with 50% and 95% probability (1 TA + 200,000 FA markers scan).

Chances to contain TA	50%		95%	
Power	No correction	Šidak	No correction	Šidak
60%	482	254	12178	10363
70%	187	41	6319	4548
80%	57	2	2712	1414
90%	9	1	752	181
95%	2	1	228	18
99%	1	1	17	1

Table 5: Block LD dependency with three true effects ($m = 3$). Entries represent the number of most significant SNPs that will contain one or more true associations with 50% and 95% probability (3 TA + 200,000 FA markers scan).

Chances to contain TA	50%		95%	
Power	No correction	Šidak	No correction	Šidak
60%	5362	3501	94932	90505
70%	2527	941	72451	61590
80%	913	90	47662	33301
90%	205	2	22001	8587
95%	44	1	9863	1662
99%	2	1	1530	14

Table 6: Two-SNP genotypic (eight degrees of freedom) tests, assuming block LD dependency between the tests. Entries are the numbers of most significant associations that will contain a dilocus true association with 50% and 95% probability (1 dilocus TA + 200,000 FA's scan).

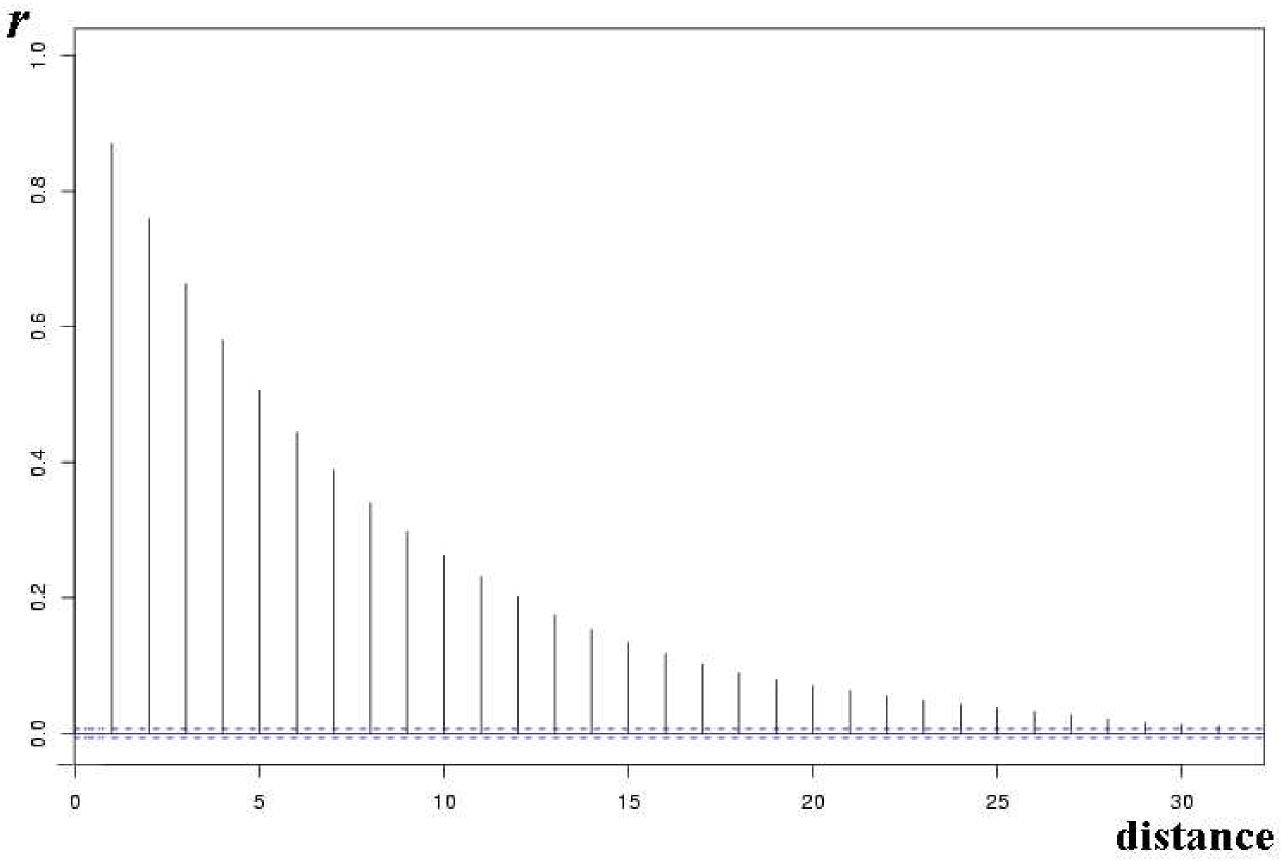

Figure 1: Diffusion-generated p-value correlation decay. The distance between two neighboring markers (one unit on the X-axis) is 15 kB.

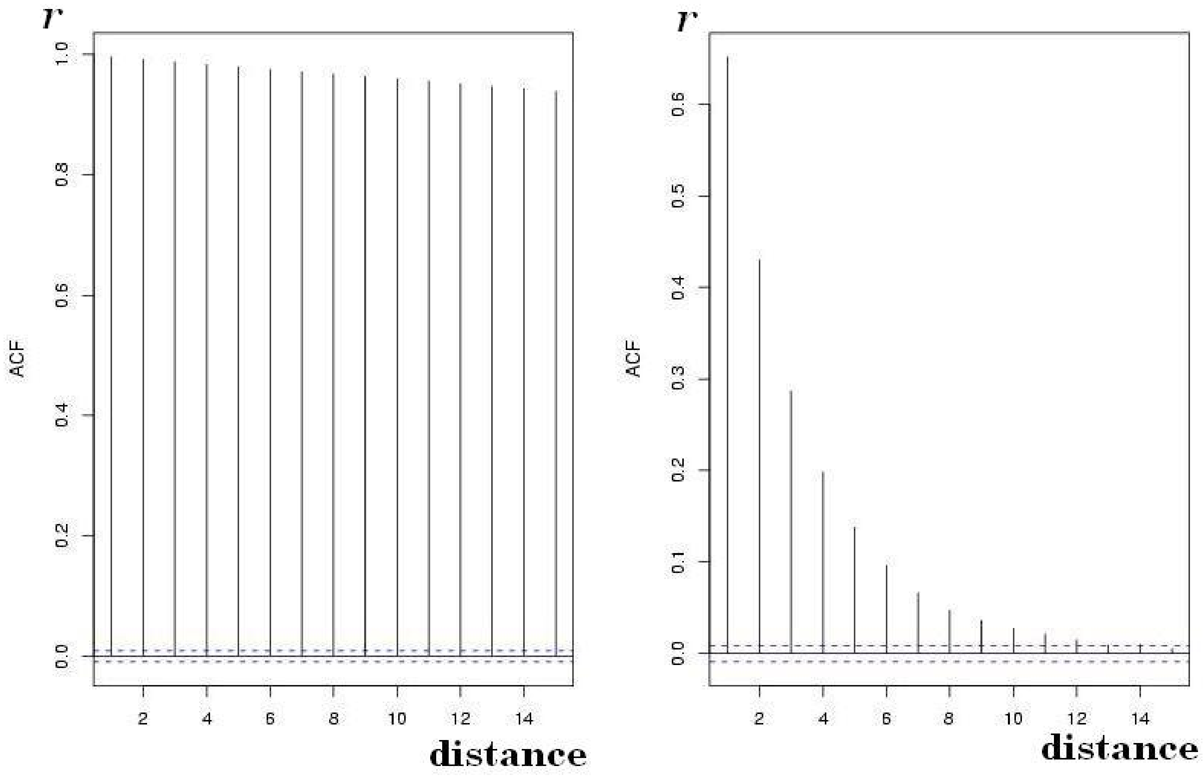

Figure 2: p-value correlation decay within (left graph) and between (right graph) LD blocks. The distance between two neighboring markers (one unit on the X-axis) is 15 kB.

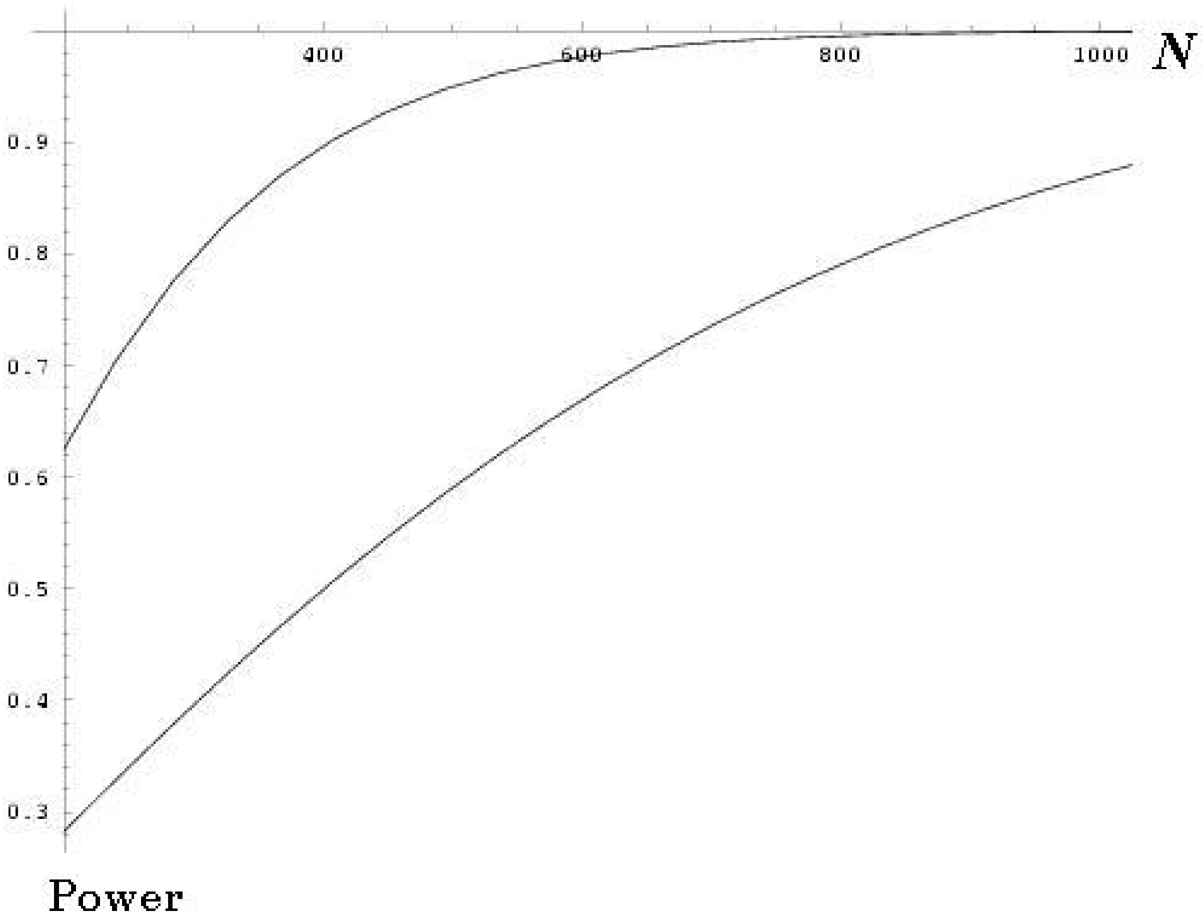

Figure 3: Expected power associated with individual effects (lower line) given overall power for three effects ($m = 3$), upper line.

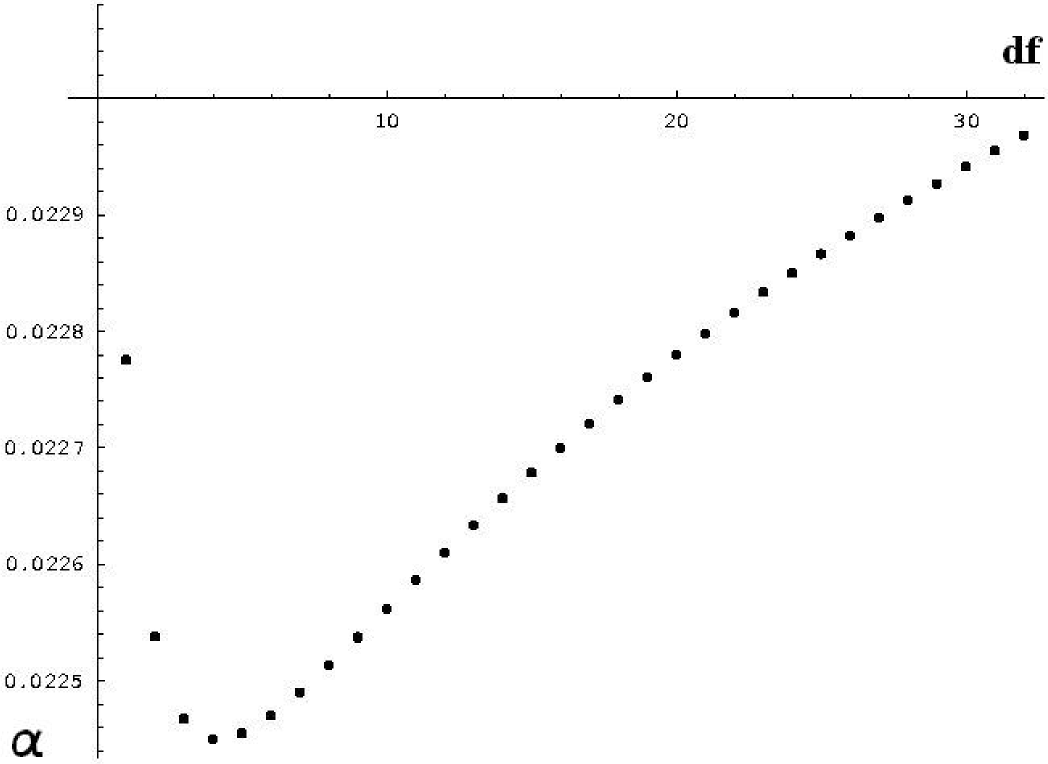

Figure 4: Plot of degrees of freedom vs. significance level (α) for 80% power tests at the 70% quantile

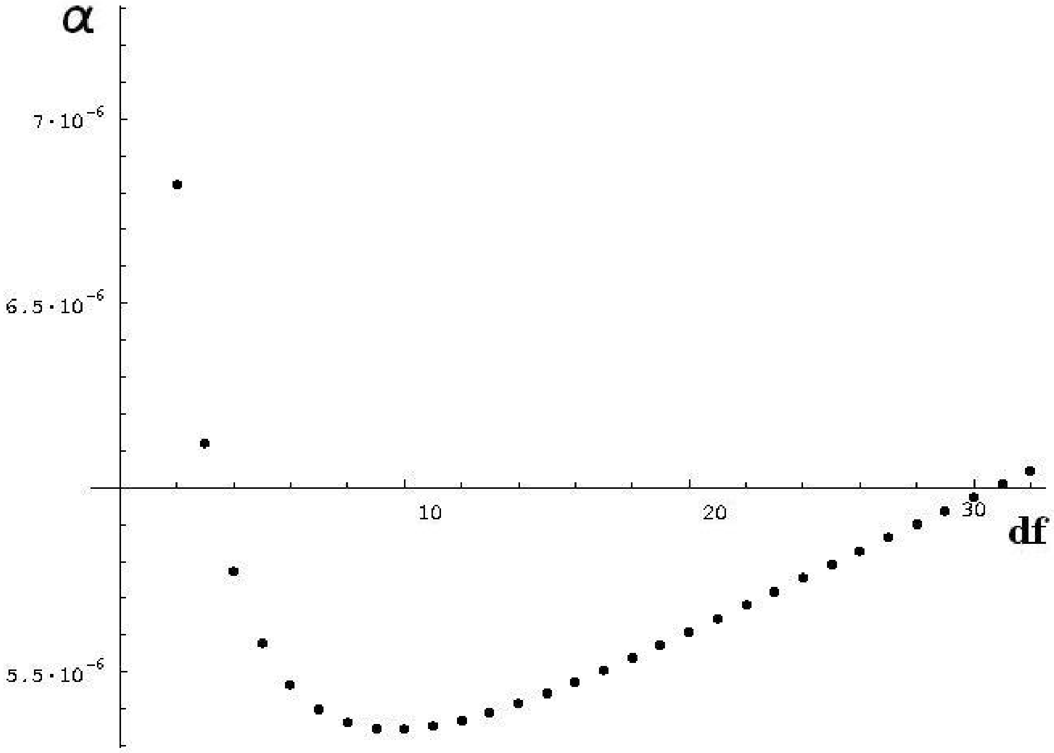

Figure 5: Plot of degrees of freedom vs. significance level (α) for 80% power tests at the 5% quantile