Report

Evaluating Statistical Significance in Two-Stage Genomewide Association Studies

D. Y. Lin

Department of Biostatistics, University of North Carolina, Chapel Hill

Genomewide association studies are being conducted to unravel the genetic etiology of complex human diseases. Because of cost constraints, these studies typically employ a two-stage design, under which a large panel of markers is examined in a subsample of subjects, and the most-promising markers are then examined in all subjects. This report describes a simple and efficient method to evaluate statistical significance for such genome studies. The proposed method, which properly accounts for the correlated nature of polymorphism data, provides accurate control of the overall false-positive rate and is substantially more powerful than the standard Bonferroni correction, especially when the markers are in strong linkage disequilibrium.

A decade ago, Risch and Merikangas (1996) suggested that genetic variants predisposing to complex human diseases could be identified through genomewide association scans involving hundreds of thousands or more markers and thousands of subjects. With the recent availability of genomewide surveys of genetic variants (The International SNP Map Working Group 2001; Hinds et al. 2005; The International HapMap Consortium 2005) and the rapid decrease in SNP genotyping costs, this vision has become a real possibility. Indeed, numerous genomewide association studies for a range of disorders are being planned or are already underway.

Despite recent advances in high-volume genotyping technology, it is still prohibitively expensive to genotype hundreds of thousands of markers in thousands of subjects. Thus, most genomewide association studies adopt a two-stage design: in the first stage, a dense set of SNP markers across the genome is genotyped and tested using a fraction of the available subjects, and, in the second stage, the most-promising markers are genotyped in the remaining subjects and tested using all subjects (Satagopan et al. 2002, 2004; Satagopan and Elston 2003; Maraganore et al. 2005; Thomas et al. 2005; Skol et al., in press).

Assessing statistical significance in such two-stage genome studies presents an important challenge. The current practice is to use the Bonferroni correction based on the total number of markers tested in the first stage (Maraganore et al. 2005; Thomas et al. 2005). This strategy is punitively conservative for two reasons. First, it assumes that none of the markers eliminated in stage 1 would reach statistical significance if they were genotyped and tested in stage 2. Second, it assumes that the test statistics are independent over all markers. The first assumption was relaxed by Skol et al. (in press). The second assumption fails when markers are in linkage disequilibrium (LD). The ENCODE data from the HapMap Project reveal that SNPs are typically in complete LD with several nearby SNPs and in strong LD with many others; thus, the Bonferroni correction is highly conservative (The International HapMap Consortium 2005).

In this report, I show how to properly incorporate the two-stage sampling and the correlation structure of the test statistics into the evaluation of statistical significance. The strategy relies on the fact that the statistics used in association testing can be represented by the so-called efficient score functions, which are sums of independent terms (see appendix A). This fact implies that the statistics are jointly normal in large samples, both over the markers and between the two stages, with correlations that can be estimated empirically from the data. I develop an efficient Monte Carlo algorithm to evaluate this joint distribution, providing appropriate thresholds for declaring statistical significance.

Suppose that a total of m markers are genotyped and tested on n_1 subjects in stage 1, and the most-promising markers are genotyped in the remaining $n_2 = n - n_1$

Received November 7, 2005; accepted for publication January 2, 2006; electronically published January 11, 2006.

Address for correspondence and reprints: Dr. Danyu Lin, Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599-7420. E-mail: lin@bios.unc.edu

Am. J. Hum. Genet. 2006;78:505–509. 0002-9297/2006/7803-0018$15.00

subjects and tested using all n subjects in stage 2. All subjects are assumed to be unrelated. For $s = 1,2$ and $j = 1, \ldots, m$, the test statistic for testing the mth marker in the sth stage can be written in the following form or can be approximated by the statistic of the following form:

$$T_j(s) = U_j(s)^T V_j(s)^{-1} U_j(s) ,$$

where

$$U_j(1) = \sum_{i=1}^{n_1} U_{ji} ,$$

$$U_j(2) = \sum_{i=1}^{n} U_{ji} ,$$

U_{ji} involves only the data from the ith subject,

$$V_j(1) = \sum_{i=1}^{n_1} U_{ji} U_{ji}^T ,$$

and

$$V_j(2) = \sum_{i=1}^{n} U_{ji} U_{ji}^T .$$

Note that U_j pertains to the efficient score function, and V_j is the covariance matrix of U_j (see appendix A). In most situations,

$$U_{ji} = (Y_i - \mu_y)(X_{ji} - \mu_j) ,$$

where Y_i is the phenotypic value of the ith subject, X_{ji} is the genotype score for the jth marker of the ith subject, and μ_y and μ_j are the population means of Y_i and X_{ji}, respectively. In the actual calculations of $T_j(s)$, μ_y and μ_j are replaced with the sample means.

Under the null hypothesis of no association, $U_j(s)$ is approximately normal with mean zero and covariance matrix $V_j(s)$ in large samples, so $T_j(s)$ has an approximate χ^2 distribution with d df, where d is the dimension of $U_j(s)$. In addition, $[U_1(1), \ldots, U_m(1), U_1(2), \ldots, U_m(2)]$ is approximately multivariate normal with mean zero and covariance matrices

$$\mathrm{Cov}[U_j(1), U_k(1)] = \mathrm{Cov}[U_j(1), U_k(2)] = \sum_{i=1}^{n_1} U_{ji} U_{ki}^T$$

and

$$\mathrm{Cov}[U_j(2), U_k(2)] = \mathrm{Cov}[U_j(1), U_k(1)] + \sum_{i=n_1+1}^{n} U_{ji} U_{ki}^T .$$

Note that $\mathrm{Cov}(Z_1, Z_2 + Z_3) = \mathrm{Cov}(Z_1, Z_2)$ if Z_1 and Z_3 are uncorrelated. The values of U_{ji} $(i = n_1 + 1, \ldots, n)$ are unknown unless the jth marker is genotyped in stage 2. However,

$$\sum_{i=n_1+1}^{n} U_{ji} U_{ki}^T$$

can be estimated by

$$\frac{n_2}{n_1} \sum_{i=1}^{n_1} U_{ji} U_{ki}^T ,$$

provided that the subjects are randomly chosen for genotyping in stage 1.

I derive a simple and efficient Monte Carlo procedure to evaluate the joint distribution of $[U_1(1), \ldots, U_m(1), U_1(2), \ldots, U_m(2)]$. Define

$$\tilde{U}_j(1) = \sum_{i=1}^{n_1} U_{ji} G_i$$

and

$$\tilde{U}_j(2) = \tilde{U}_j(1) + \left(\frac{n_2}{n_1}\right)^{\frac{1}{2}} \sum_{i=1}^{n_1} U_{ji} G_i' ,$$

where $G_1, \ldots, G_{n_1}, G_1', \ldots, G_{n_1}'$ are independent standard normal random variables. Also, define

$$\tilde{T}_j(s) = \tilde{U}_j(s)^T V_j(s)^{-1} \tilde{U}_j(s) .$$

Conditional on the observed data, $[\tilde{U}_1(1), \ldots, \tilde{U}_m(1), \tilde{U}_1(2), \ldots, \tilde{U}_m(2)]$ is multivariate normal with mean zero and (approximately) the same covariance matrix as $[U_1(1), \ldots, U_m(1), U_1(2), \ldots, U_m(2)]$. Thus, one can use the joint distribution of $[\tilde{T}_1(1), \ldots, \tilde{T}_m(1), \tilde{T}_1(2), \ldots, \tilde{T}_m(2)]$ to approximate that of $[T_1(1), \ldots, T_m(1), T_1(2), \ldots, T_m(2)]$.

Suppose that the jth marker is selected for genotyping in stage 2 if $T_j(1) > c_1$, where c_1 is chosen to achieve a certain level of statistical significance or to yield a desired proportion of markers for stage 2 testing. The null hypothesis of no association between the jth marker and disease is rejected if $T_j(1) > c_1$ and $T_j(2) > c_2$, where c_2 is chosen so that, under the global null hypothesis of no association,

$$\Pr[T_j(1) > c_1 \text{ and } T_j(2) > c_2 \text{ for some } j] = \alpha ,$$

where α is the nominal type I error rate or significance level. One can approximate this equation by

$$\Pr[\tilde{T}_j(1) > c_1 \text{ and } \tilde{T}_j(2) > c_2 \text{ for some } j] = \alpha . \quad (1)$$

The probability on the left-hand side of equation (1) is estimated by generating a large number, say 10,000, of realizations of $\tilde{T}_j(1)$ and $\tilde{T}_j(2)$.

Given c_1 and α, one can use equation (1) to determine c_2. This calculation can be done through a bisection search based on a single set of realizations of $\tilde{T}_j(1)$ and $\tilde{T}_j(2)$. In practice, c_2 on the left-hand side of equation (1) is replaced with the observed value of $T_j(2)$, and significant association with the jth marker is declared if the resulting probability is $<\alpha$.

To assess the performance of the proposed method, I simulated 10,000 SNPs with minor-allele frequencies of 0.3 and varying degrees of LD. I set the disease prevalence in the population to be ~5%. Under the null hypothesis, none of the SNP markers was associated with disease. Under the alternative hypothesis, the minor allele of SNP 5,000 had a dominant effect with a relative risk of 1.5. I selected 1,000 cases and 1,000 controls and used the Pearson χ^2 statistic under the dominant model to test the association between each SNP and disease status. I set the nominal significance level at 0.05.

Figure 1 displays the results for the two-stage design, under which 50% of the cases and controls are genotyped in stage 1; c_1 was set at 3, so that ~10% of the markers are selected for genotyping in stage 2. The results for other designs are similar and thus omitted. The empirical type I error rate pertains to the probability of finding any association under the null hypothesis, and the empirical power pertains to the probability of identifying SNP 5,000 under the alternative hypothesis. Each of these probabilities was estimated from 1,000 simulated data sets; for each data set, the Monte Carlo evaluation was based on 10,000 normal samples.

As shown in figure 1, the proposed method maintains its type I error near the nominal level, whereas the Bonferroni correction is conservative. The type I error rates of the proposed method are ~0.052, 0.055, and 0.050 when the squared correlation coefficient, r^2, between two adjacent markers is 0.5, 0.9, and 0.99, respectively. By contrast, the corresponding type I rates based on the Bonferroni correction are ~0.037, 0.022, and 0.002.

The proposed method is considerably more powerful than the Bonferroni correction, especially when the markers are in strong LD. The power of the proposed method is ~75%, whereas that of the Bonferroni correction is ~65%, when r^2 between two adjacent markers is 0.9; the corresponding power estimates are 85% and 60% when r^2 is 0.99. For the Bonferroni correction to achieve the same power as that of the proposed method, the sample sizes would need to be increased by ~15% and 30% when the r^2 values between two adjacent markers are 0.9 and 0.99, respectively. Thus, the power advantages of the proposed method have important implications.

In the studies above, LD was created by allowing the

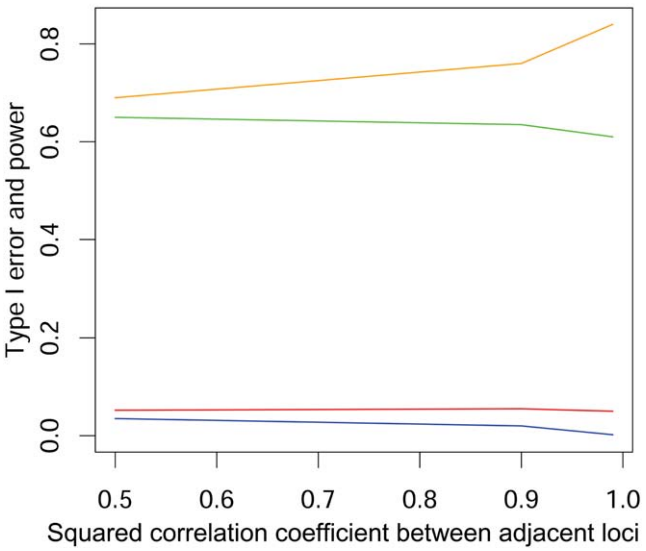

Figure 1 Empirical type I error rate and power at the nominal significance level of 0.05. The red and orange curves correspond to the type I error and power of the proposed method, respectively, and the blue and green curves correspond to the type I error and power of the Bonferroni correction, respectively. The *X*-axis pertains to the squared correlation coefficient, r^2, between two adjacent markers, which varies from 0.5 to 0.99.

SNP allele frequencies of each marker to depend on those of the preceding marker, so that the LD decays exponentially as the intermarker distance increases. When r^2 is 0.9 between two adjacent markers, the value of r^2 is 0.8 between every second marker and 0.3 between SNPs that are 10 markers apart; when r^2 is 0.99 between two adjacent markers, the value of r^2 is 0.98 between every second marker and 0.8 between SNPs that are 20 markers apart.

To generate more-realistic LD structures, I considered the HapMap data (The International HapMap Consortium 2005). I simulated data in the same manner as in the studies above, except that the genotypes were sampled from the white phasing data in the ENCODE region of chromosome 4, which consists of 1,393 SNPs. Under the null hypothesis of no association, the type I error rates were found to be ~0.047 and 0.009 for the proposed and Bonferroni methods, respectively. Under the alternative hypothesis that SNP 1,100, which has a minor-allele frequency of ~0.2, had a dominant effect with a relative risk of 1.5, the power was ~0.91 for the proposed method, compared with 0.78 for the Bonferroni correction.

The results in figure 1 pertain to 10,000 SNPs, which is approximately the number of markers on a single chromosome in the currently available 100K–500K SNP platforms. Since the test statistics are generally uncorrelated among the chromosomes, the proposed method

can be applied to each chromosome separately. It is unclear whether one should adjust for multiple comparisons among chromosomes when hundreds of thousands or more markers are tested. It is perhaps more sensible to control a few (say, 10–20) false positives rather than a single one in such massive-scale hypothesis testing (Lehmann and Romano 2005).

The above studies were concerned with single-locus effects. The proposed method is certainly applicable to multilocus searches, including interactions and haplotype effects (Epstein and Satten 2003). The method is also potentially useful for complex multistage studies.

This method combines the raw data from the two stages in the final analysis. An alternative approach is to combine the two test statistics (i.e., to sum the two standardized statistics) (Skol et al., in press). It is trivial to modify this method for the combined test statistics, provided that the subjects are randomly selected for genotyping in stage 1. However, a major motivation for combining the two test statistics is to allow for heterogeneity between the first-stage and second-stage samples. The work of Zaykin et al. (2002) and Dudbridge and Koeleman (2004) can also be extended to two-stage studies through this Monte Carlo approach. Although I have focused on studies of unrelated individuals, the proposed method can be adapted to family studies by changing "subject" to "family" in the description.

Because of the two-stage sampling, the method described here is different from that of Lin (2005*a*, 2005*b*). In particular, the new Monte Carlo procedure circumvents the problem that the genotype data are unobserved for those markers eliminated in the first stage.

Unlike for single-stage studies, it is not possible to evaluate statistical significance for two-stage studies by permutation. If the value of $T_j(1)$ based on the original data does not exceed c_1, then the jth marker is not genotyped in stage 2. When the data are permuted, the value of $T_j(1)$ based on the permuted data may exceed c_1. In that case, one needs to evaluate $T_j(2)$ on the basis of the permuted data, but that evaluation is not possible because the jth marker is missing in all n_2 subjects.

The proposed method provides an essential ingredient for designing genomewide association studies. In the design stage, one would simulate the genotype data for the specific SNPs to be tested and use equation (1) to determine c_2. One would then determine the power by evaluating the probabilities of true detection for various relative risks through simulation.

I found that two-stage designs in which ~50% of the available subjects are genotyped in stage 1 and the top 1%–10% of the markers are genotyped in stage 2 are nearly as powerful as the single-stage design that genotypes all markers in all subjects (data not shown). Similar findings were reported elsewhere for independent test statistics (Satagopan and Elston 2003; Skol et al., in press). Thus, two-stage designs are highly cost effective. With the Bonferroni correction, the penalty is proportional to the number of markers tested in stage 1, regardless of the marker density. By contrast, the proposed method properly accounts for the actual correlations of SNPs and does not unfairly penalize SNP platforms with very high density.

A computer program that implements the proposed method is freely available at the author's Web site (see Web Resource section). The computing time is linear in relation to the number of markers and the number of subjects. The analysis for a typical genome scan (100K–500K markers and 1,000–5,000 subjects) can be completed in a short amount of time on any high-performance computer.

Acknowledgments

This research was supported by National Institutes of Health grants 2 R37 GM047845-15 and 2 R01 CA082659-08. The author thanks Drs. Michael Boehnke, Fred Wright, and Donglin Zeng for helpful discussions.

Appendix A

Score Statistics

For quantitative traits, it is natural to consider the linear regression model

$$Y_i = \mu + \beta^T X_{ji} + \epsilon_i \,, \tag{A1}$$

where X_{ji} is the ith subject's genotype score for the jth marker and ϵ_i is normal with mean 0 and variance σ^2. For simplicity of description, the dependence of the parameters on j is suppressed. Under the additive model, X_{ji} denotes the number of minor alleles that the ith subject has; under the dominant (or recessive) model, X_{ji} indicates, with values of 1 and 0, whether or not the ith subject has at least one minor allele (or, for the recessive model, two minor alleles); under the codominant model, X_{ji} consists of two components indicating one and two minor alleles. For dichotomous traits, it is common to employ the logistic regression model

$$P(Y_i = 1) = \frac{e^{\nu+\beta^T X_{ji}}}{1 + e^{\nu+\beta^T X_{ji}}} \,. \tag{A2}$$

One is interested in testing the null hypothesis H_0: $\beta = 0$ against the alternative hypothesis H_1:$\beta \neq 0$ at every marker. The parameters μ and σ^2 in model (A1) and ν in model (A2) are nuisance parameters, which are denoted by η. There are three asymptotically equivalent test statistics: the Wald statistic, the likelihood-ratio sta-

tistic, and the score statistic. Here, it is convenient to work with score statistics.

The log-likelihood function for (β,η) at the jth marker is $l_j(\beta,\eta) = \sum_{i=1}^{n} l_{ji}(\beta,\eta)$, where $l_{ji}(\beta,\eta)$ pertains to the contribution from the ith subject. Let $U_{\beta,ji}(\beta,\eta) = \partial l_{ji}(\beta,\eta)/\partial\beta$ and $U_{\eta,ji}(\beta,\eta) = \partial l_{ji}(\beta,\eta)/\partial\eta$. The score statistic for testing $H_0{:}\beta = 0$ takes the form

$$U_j = \sum_{i=1}^{n} U_{\beta,ji}(0,\tilde{\eta}) \ , \qquad \text{(A3)}$$

where $\tilde{\eta}$ is the (restricted) maximum-likelihood estimator of η under H_0—that is, the solution to the equation $\sum_{i=1}^{n} U_{\eta,ji}(0,\eta) = 0$. Note that U_j is the score function for β evaluated at $\beta = 0$ and $\eta = \tilde{\eta}$ and is not a sum of independent terms for a given j. It follows from the Taylor series expansions and the law of large numbers that $n^{-1/2}U_j$ has the same asymptotic distribution as $n^{-1/2}\sum_{i=1}^{n} U_{ji}$, where

$$U_{ji} = U_{\beta,ji}(0,\eta) - \Sigma_{\beta\eta}(0,\eta)\Sigma_{\eta\eta}^{-1}(0,\eta)U_{\eta,ji}(0,\eta)$$

and $\Sigma_{\beta\eta}(\beta,\eta)$ and $\Sigma_{\eta\eta}(\beta,\eta)$ are the limits of $n^{-1}\partial^2 l_j(\beta,\eta)/\partial\beta\partial\eta$ and $n^{-1}\partial^2 l_j(\beta,\eta)/\partial\eta^2$ as n goes to infinity (Cox and Hinkley 1974, section 9.3(iii)). One calls U_{ji} the ith subject's efficient score function. Under both models (A1) and (A2),

$$U_{ji} = (Y_i - \mu_y)(X_{ji} - \mu_j) \ ,$$

where μ_y and μ_j are the population means of Y_i and X_{ji}, respectively. Since U_{ji} involves only the observations from the ith subject, U_{ji} are independent zero-mean random vectors for any given j. Thus, it follows from the multivariate central limit theorem that, under the null hypothesis of no association, $n^{-1/2}(U_1,\ldots,U_m)$ is asymptotically multivariate normal with mean 0 and with $n^{-1}\sum_{i=1}^{n} U_{ji}U_{ki}^T$ as the covariance matrix between the jth and kth markers.

In the actual calculations of the test statistics, the unknown parameters in U_{ji} are replaced with the (restricted) maximum-likelihood estimators. Since $\sum_{i=1}^{n} U_{\eta,ji}(0,\tilde{\eta}) = 0$ by the definition of $\tilde{\eta}$, the replacement of η with $\tilde{\eta}$ in U_{ji} yields $\sum_{i=1}^{n} U_{ji} = \sum_{i=1}^{n} U_{\beta,ji}(0,\tilde{\eta})$, which is consistent with the definition of the score statistic given in equation (A3). It can be shown that, under model (A1) with a dichotomous genotype score,

$$\frac{\sum_{i=1}^{n} U_{ji}}{\left(\sum_{i=1}^{n} U_{ji}^2\right)^{\frac{1}{2}}} = \frac{\overline{Y}_1 - \overline{Y}_2}{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^{\frac{1}{2}}} \ ,$$

where n_1 and n_2 are the numbers of subjects in the two groups and $(\overline{Y}_1,\overline{Y}_2)$ and (S_1^2,S_2^2) are the sample means and sample variances in the two groups. This is, of course, the well-known two-sample t statistic. Likewise, the familiar Pearson χ^2 statistics can be generated under model (A2).

The above description pertains to single-stage studies. However, all the results can be extended to two-stage designs in an obvious manner.

Web Resource

The URL for data presented herein is as follows:

Author's Web site, http://www.bios.unc.edu/~lin/

References

Cox DR, Hinkley DV (1974) Theoretical statistics. Chapman and Hall, New York

Dudbridge F, Koeleman BPC (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

Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73: 1316–1329

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079

Lehmann EL, Romano JP (2005) Generalizations of the familywise error rate. Ann Stat 33:1138–1154

Lin DY (2005*a*) An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 21:781–787

——— (2005*b*) On rapid simulation of *P* values in association studies. Am J Hum Genet 77:513–514

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PVK, Frazer KA, Cox DR, Ballinger DG (2005) High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 77:685–693

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

Satagopan JM, Elston RC (2003) Optimal two-stage genotyping in population-based association studies. Genet Epidemiol 25:149–157

Satagopan JM, Venkatraman ES, Begg CB (2004) Two-stage designs for gene-disease association studies with sample size constraints. Biometrics 60:589–597

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

Skol AD, Scott LJ, Abecasis GR, Boehnke M. Forget replication: joint analysis is more efficient for genomewide association studies. Nat Genet (in press)

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Thomas DC, Haile RW, Duggan D (2005) Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet 77:337–345

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