BIOMETRICS 55, 997–1004
December 1999

Genomic Control for Association Studies

B. Devlin

Department of Psychiatry, University of Pittsburgh,
3811 O'Hara Street, Pittsburgh, Pennsylvania 15213
email: devlinbj@msx.upmc.edu

and

Kathryn Roeder

Department of Statistics, Carnegie Mellon University,
5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213
email: roeder@stat.cmu.edu

SUMMARY. A dense set of single nucleotide polymorphisms (SNP) covering the genome and an efficient method to assess SNP genotypes are expected to be available in the near future. An outstanding question is how to use these technologies efficiently to identify genes affecting liability to complex disorders. To achieve this goal, we propose a statistical method that has several optimal properties: It can be used with case–control data and yet, like family-based designs, controls for population heterogeneity; it is insensitive to the usual violations of model assumptions, such as cases failing to be strictly independent; and, by using Bayesian outlier methods, it circumvents the need for Bonferroni correction for multiple tests, leading to better performance in many settings while still constraining risk for false positives. The performance of our genomic control method is quite good for plausible effects of liability genes, which bodes well for future genetic analyses of complex disorders.

KEY WORDS: Bayesian inference; Case–control; Complex genetic disorder; Outliers; Population heterogeneity; Single nucleotide polymorphism genotypes.

1. Introduction

A spin-off of the Human Genome Project is the massive governmental and industry-sponsored effort to develop a dense set of biallelic markers (single nucleotide polymorphisms; SNP) throughout the human genome (Collins et al., 1998; Wang et al., 1998). Coupled with this effort is intense research to produce technology to assess SNP genotypes rapidly and economically. These efforts have been spurred by the realization that a dense set of SNP throughout the genome could yield critical information for determining the genetic basis of complex diseases (Risch and Merikangas, 1996), in large part through population-level association induced by the interplay of linkage and evolution.

An outstanding question is how to use SNP technology efficiently. One possibility is to apply it to case–control samples. Case–control studies have numerous advantages for the genetic dissection of complex traits (Morton and Collins, 1998; Risch and Teng, 1998). Case–control studies have been criticized, however, because they rely on the unrealistic assumption of population homogeneity; in the face of population heterogeneity, spurious associations can arise (Li, 1972). Therefore, alternative methods, which employ family-based sampling to obviate the effects of population heterogeneity (Falk and Rubinstein, 1987; Spielman, McGinnis, and Ewens, 1993; Curtis, 1997), have become increasingly popular.

Despite population heterogeneity, case–control designs are appealing because they do not require recruitment of additional family members for cases, which can be expensive at best. What is needed is a method that has the advantages of both case–control and family-based designs. In this article, we propose such a method for either SNP association scans or tests of candidate genes. For case–control data, our method effectively uses the genome itself to induce controls similar to family-based studies and to determine what constitutes a significant departure from the null model of no linkage disequilibrium.

An advantage of dense association genomic scans is that they can detect loci having a small impact on risk to human disorders (Risch and Merikangas, 1996). A disadvantage is that a large number of false positives occur when many significance tests are conducted. A traditional solution is to impose Bonferroni correction. Instead we propose a Bayesian outlier test as a means of determining which markers exhibit significant linkage disequilibrium with the disorder. In essence, the outlier test bypasses the usual rigid assumptions required to obtain chi-square distributed random variables in favor of more flexible statistics and weaker assumptions. This test is appropriate for family-based and case–control designs. For this article, however, we focus on the latter.

Another feature of our proposed methodology is that it allows for violations in the usual model assumption, independence of observations, which, when violated, leads to extra variance in the test statistic. For instance, for case–control studies, affected individuals are more likely to be related than are control individuals because they share a genetic disorder and, ideally, a common genetic basis for the disorder. In fact, this is the *sine qua non* of association-based genetic studies. Hence, for case–control studies, test statistics are generally inflated relative to expectation under the assumption of an independent sample and no genetic association with the disease. For this reason, simple marker-by-marker hypothesis tests will almost surely produce false positives, even after a Bonferroni correction. These false positives often are attributed to population heterogeneity, but we offer cryptic relatedness as a more important explanation.

Our proposed method, when applied to case–control studies, does not require knowledge of the genealogy of the population or the nature of population heterogeneity. The test adapts and corrects for problems arising from population heterogeneity, poor choice of controls, and cryptic relatedness of cases, albeit at a cost in power. Our goal in this article is to describe the method and assess its power for reasonable choices of population and genomic characteristics.

2. Methods

2.1 *The Data and Genetic Models*

2.1.1. *Properties of a single locus.* For a case–control study and n biallelic markers, the data for each marker are given in a standard 2×3 table of genotype by case and control (see Table 1). To test for lack of independence, three 1-d.f. chi-square statistics are possible, corresponding to dominant, recessive, and additive genetic models. For an association genome scan to assess the genetic basis of a complex disorder, there is usually no prior information about mode of inheritance. In this setting, then, an additive model should perform well, and this is the model that we will investigate in depth. The additive genetic model can be tested using Armitage's trend test (Armitage, 1955),

$$Y^2 = \frac{N\{N(r_1 + 2r_2) - R(n_1 + 2n_2)\}^2}{R(N-R)\{N(n_1 + 4n_2) - (n_1 + 2n_2)^2\}}. \quad (1)$$

This test is equivalent to the score test in the logistic regression model.

For each marker, the data also can be summarized via a 2×2 allelic table (Table 2). (See Sasieni [1997] for a thorough analysis of the features of allelic versus genotypic analyses.) Here we review some of his results and explore these issues further, as they are critical to our methodological development.

Based on the allelic data, the chi-square test for association is

$$Y_A^2 = \frac{2N\{2N(r_1 + 2r_2) - 2R(n_1 + 2n_2)\}^2}{(2R)2(N-R)\{2N(n_1 + 2n_2) - (n_1 + 2n_2)^2\}}. \quad (2)$$

Table 1
Genotype distribution

	A_1 alleles			
	0	1	2	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

The numerators of both trend and allelic tests are proportional to the square of the weighted difference between the number of A_1 alleles in the cases and the controls, $N(r_1 + 2r_2) - R(n_1 + 2n_2) = S(r_1 + 2r_2) - R(s_1 + 2s_2)$. The tests differ due to their denominators and, as we shall see shortly, by their assumptions concerning independence, i.e.,

$$Y_A^2/Y^2 = 1 + 4n_0n_2 - n_1^2/\{(n_1 + 2n_2)(n_1 + 2n_0)\}. \quad (3)$$

Under independence or Hardy–Weinberg equilibrium in the population, this ratio is approximately equal to one (Sasieni, 1997). The trend test is more conservative than the allelic test because this ratio tends to be greater than one, even under the null hypothesis, when the population is not in Hardy–Weinberg equilibrium.

Let F denote Wright's coefficient of inbreeding (Elandt-Johnson, 1971, p. 214). Here we define inbreeding broadly to denote any form of mating leading to increased homozygosity. Two distinct processes lead to this end: matings among relatives and population substructure. For either process, F measures the correlation between uniting gametes, but the value and meaning of F is context dependent.

If p_i is the frequency of A_i in the population, then the genotypic frequencies are described by

$$\Pr(A_iA_j) = \begin{cases} Fp_i + (1-F)p_i^2 & \text{if } i = j \\ 2(1-F)p_ip_j & \text{if } i < j. \end{cases} \quad (4)$$

Let G be the number of A_1 alleles in the genotype of a single individual. Clearly, E$[G]$ is not a function of F, but the variance of G is inflated by a factor of $(1+F)$ from that expected for a population in Hardy–Weinberg equilibrium, var$[G] = (1+F)2p_1p_2$ (cf., Elandt-Johnson, 1971, pp. 216–218).

Table 2
Allele distribution

	A_1	A_2	Total
Case	$r_1 + 2r_2$	$r_1 + 2r_0$	$2R$
Control	$s_1 + 2s_2$	$s_1 + 2s_0$	$2S$
Total	$n_1 + 2n_2$	$n_1 + 2n_0$	$2N$

Noting that $\text{E}(n_1 + 2n_2)/2N = p_1$ and replacing $n_i, i = 0, 1, 2$, by their expected values reveals that (3) is approximately equal to $(1+F)$ under any population genetic model for which genotype probabilities are given by (4). Consequently, we see that the trend test automatically accounts for the extra-binomial variance induced by correlation of uniting gametes. What of correlation across individuals? The trend test assumes that the genotypes of individuals are independent, but that assumption is false if there is population substructure or related individuals within one or both of the samples.

In fact, concern about the effect of population substructure on case–control studies is common (Spielman et al., 1993). The Wahlund effect, a well-known result of substructure, predicts an allelic correlation within genotypes, which results in an excess of homozygotes in a substructured population. As we just saw, the trend test accounts for this effect. More troublesome, then, is the fact that the allelic correlation extends across individuals within the subpopulation as well. For a substructured population, F is also the correlation between alleles from members of the same subpopulation. As a consequence of this correlation, the usual chi-square analyses can result in a rate of false positives exceeding the nominal level. As we demonstrate below, whether or not excess false positives occur depends on the nature of the substructure.

Let G_i, $i = 1, \ldots, R$, denote the number of A_1 alleles in the ith case. Let H_j, $j = 1, \ldots, S$, denote the same for the controls. Let $a_1, a_2, \ldots, a_m$ and $b_1, b_2, \ldots, b_m$ denote the sample size of cases and controls from each of the m subpopulations, $\Sigma\, a_k = R$ and $\Sigma\, b_k = S$. For simplicity of exposition, take $R = S$. The trend and allelic tests are proportional to the square of $T = \Sigma_i\, G_i - \Sigma_j\, H_j$. We analyze the behavior of the test statistic T under the null hypothesis. The variance of T is highly dependent on the similarity between a_k and b_k,

$$\begin{aligned}\operatorname{var}(T) &= \sum_{i=1}^{R} \operatorname{var}(G_i) + \sum_{j=1}^{S} \operatorname{var}(H_j)\\ &\quad + 2\sum_{i<l} \operatorname{cov}(G_i, G_l) + 2\sum_{j<l} \operatorname{cov}(H_j, H_l)\\ &\quad - 2\sum_{i}\sum_{j} \operatorname{cov}(G_i, H_j).\end{aligned}$$

From the above, we have $\operatorname{var}(G_i) = \operatorname{var}(H_j) = 2p_1p_2(1+F)$. For any pair of genotypes from the same subpopulation, $\operatorname{cov}(G_i, G_l) = \operatorname{cov}(H_j, H_l) = \operatorname{cov}(G_i, H_j) = 4Fp_1p_2, i \neq l, j \neq l$.

It follows that the variance of the difference above equals

$$\begin{aligned}&4Rp_1p_2(1+F)\\ &\quad + 4Fp_1p_2 \sum_k \{a_k(a_k-1) + b_k(b_k-1) - 2a_kb_k\}. \quad (5)\end{aligned}$$

This quantity achieves its maximum, $2Rp_1p_2(2+F(2R-1))$, when a_i takes the value R for some i and $a_j = 0$, $j \neq i$, and b_k takes the value S for some k, $k \neq i$, and $b_j = 0$, $j \neq k$. Its minimum, $4Rp_1p_2(1-F)$, occurs when $a_k = b_k, k = 1, \ldots, m$. Contrast these with the limiting variance utilized in the trend test, $4Rp_1p_2(1+F)$, and the allelic test, $4Rp_1p_2$. Define $\lambda = \operatorname{var}(T)/\{4Rp_1p_2(1+F)\}$ as the variance inflation factor relative to the trend test.

The most extreme effect of substructure occurs if cases and controls define distinct subpopulations. In this instance, even small values of F can have a large impact on the variance of T. Alternatively, it is optimal for affection status to be independent of subpopulation membership. In this scenario, population substructure has essentially no impact on the distribution of T. In fact, at its minimum, $\lambda = (1-F)/(1+F)$. For most cases, however, the probability of affection varies somewhat by subpopulation. To see its effect, take $F = 0.05$, $R = S = 100$, $m = 10$, and $a_k = b_l = 16$ for $k = 1, \ldots, 5$ and $l = 6, \ldots, 10$ and $a_k = b_l = 4$ for $k = 6, \ldots, 10$ and $l = 1, \ldots, 5$. This fairly extreme scenario results in a variance inflation factor λ of 1.3. For a more realistic level of admixture, $F = 0.01$, λ is only 1.06.

In a case–control study of a disease with a genetic basis, cases are likely to be related; after all, they share a genetic disorder. By contrast, the controls are more likely to be independent, but they too may be related to a minor degree. For an inbred population, F is the probability uniting gametes that are identical by descent (i.b.d.). The kinship coefficient gives a related quantity: For relatives i and j, it is the probability that an allele selected randomly from i and an allele selected randomly from the same autosomal gene of j are i.b.d. In both cases, F can be interpreted as the correlation between alleles.

Because cryptic relatedness among affected individuals may have a large impact on a case–control study, we turn our attention to this case. For simplicity, consider a case–control sample with $R = S$; an allelic correlation equal to F_1 (F_2) is assumed for all individuals in the case (control) sample. Case and control samples are independent. This model is mathematically equivalent to assuming the most extreme substructure except that F_1 need not equal F_2. Under this model, $\operatorname{var}[\Sigma_i\, G_i] = 2Rp_1p_2 \times \{1 + F_1(2R-1)\}$. A similar argument holds for the controls. Consequently, under the null hypothesis of no genetic association,

$$\operatorname{var}[T] = 2Rp_1p_2 \times \{2 + (F_1 + F_2) \times (2R-1)\}. \quad (6)$$

Thus, even for small values of F_1 and F_2, the variance of T is substantially inflated over the binomial variance and it increases as a function of the sample size. For example, if $F_1 = 0.001$, $F_2 = 0$, and $R = S = 1000$, λ is 2; with $R = S = 2000$, λ is 3.

Compare this result with the admixture example given above for $F = 0.01$. Consider a sample of cases who are cryptically related and assume the case and control subpopulations differ somewhat with $F_1 = 0.0075$, $F_2 = 0.0025$, and $F_1 + F_2 = 0.01$. For $R = S = 100$, the variance inflation is 2 versus 1.06 for substructure alone. While these examples are quite artificial, the same arguments apply for more complicated instances of cryptic relatedness because the variance of T is the product of the binomial variance and a complex function of various kinship coefficients.

2.1.2. *Variance inflation estimated from multiple markers.* With data from a single locus, it is impossible to correct for the effect of population substructure and cryptic relatedness. This fact motivated development of matched case–control designs in epidemiology generally and family-based designs for genetic epidemiology. When a set of SNP is evaluated for cases and controls, however, it is possible to simultaneously estimate the variance inflation and adjust the test for association of each locus with the disorder.

In the ideal case, the inflation factor λ would be a constant for all markers. For the model of cryptic relatedness, the variance inflation is due to correlations or kinship coefficients unrelated to properties of individual loci, and thus it is the same for all markers throughout the genome. For the variance inflation due to locus-specific attributes, the results are not as transparent, but under certain conditions, the inflation factor is roughly constant. Several conditions must be met: (a) the loci under study must not have very different mutation rates (Chakraborty and Jin, 1992); (b) they cannot be under strong and subpopulation-specific selection (Crow and Kimura, 1970); and (c) with respect to population substructure, F should not vary greatly across loci. The advantage of SNP for our analyses is that they are assumed to have a minuscule mutation rate, thus meeting condition (a). Little is known about selection on the human genome, but strong, differential selection for extant SNP alleles seems unlikely. Thus, it is plausible that condition (b) is met. At issue is whether or not F varies greatly across loci.

According to Lewontin and Krakauer (1973), the variance in F is negligible, provided the number of subpopulations is large and F is small. Under a more complex model of relationships among subpopulations, Robertson (1975) derived a different expression that allows for the possibility of considerably more variance in F across the genome. Lewontin and Krakauer's results are based on a model that assumes that all subpopulations are equally related. Robertson's model more nearly describes the world's subpopulations because subpopulations within an ethnic group/race are more closely related than subpopulations across ethnic groups (e.g., Devlin, Risch, and Roeder, 1993).

In a well-designed case–control study, subjects are drawn from the same ethnic group or additional heterogeneity is modeled explicitly. Take, e.g., a random sample of Caucasians drawn from Europe. Some rough calculations based on the results of Cavalli-Sforza, Menozzi, and Piazza (1994) for 122 classical genetic markers suggest an average $F = 0.0006$ and standard deviation 0.0012. Clearly, F is not constant, but, as Lewonton and Krakauer predicted, its variance is not large for such a sample. For the remainder of the methodological development, we will assume λ is constant across loci. The impact of variation in λ will be described in the sequel.

2.2 *Statistical Analysis*

To determine which markers are in association with the disorder, we first propose a Bayesian outlier model that automatically corrects for violations of the independence model. The model uses the results for a set of loci (e.g., a genome scan) to estimate the variance inflation, λ. Formally, the Bayesian framework of this model is similar to the one proposed by Verdinelli and Wasserman (1991) for general outlier detection. A less flexible, but simpler, frequentist solution is described at the close of this section.

2.2.1. *The Bayesian approach.* For marker locus i, we obtain a statistic Y_i^2 using the trend test, $i = 1, \ldots, n$. For this report, we assume the statistics are independent (see Section 4 for further discussion). When the marker is in linkage equilibrium with the disorder and there is no population substructure or cryptic relatedness, Y_i^2 is distributed as $\mathcal{X}_1^2(0)$. We expand this null model to allow for extra variance by assuming $Y_i^2/\lambda \sim \mathcal{X}_1^2(0)$.

To allow for outliers (i.e., markers associated with the disorder), the model is enhanced so that the distribution for Y_i^2 is a mixture of chi-square distributions, i.e.,

$$Y_i^2/\lambda \sim \epsilon\, \mathcal{X}_1^2(A_i^2) + (1-\epsilon)\, \mathcal{X}_1^2(0), \qquad (7)$$

where ϵ is the prior probability a given observation is an outlier and A_i^2 is the noncentrality parameter associated with the ith outlier. It follows that $Y_i \sim \epsilon \mathrm{N}(A_i, \lambda) + (1-\epsilon)\mathrm{N}(0, \lambda)$. To simplify computations, an auxiliary variable δ_i is introduced, $\delta_i \sim \text{Bernoulli}(\epsilon)$. Given δ_i, $Y_i \sim \mathrm{N}(\delta_i A_i, \lambda)$.

We observe $X_i = |Y_i|$, not Y_i. The latter would be observable only if we knew *a priori* which allele was potentially associated with the disorder. When $\delta_i = 0$, knowing X_i is sufficient for inferential purposes. When $\delta_i = 1$, we assume $Y_i = X_i$. If $A_i/\lambda^{1/2} > 2$, then $\Pr(Y_i > 0)$ is high. When $A_i/\lambda^{1/2}$ is substantially less than two, it is not possible to distinguish this observation from the null model. Thus δ_i is taken to be zero with high probability, and we incur little error with our approximation. Finally, because the vast majority of the markers are not associated with the disorder of interest, this approximation has little effect on our inferences.

To complete our Bayesian probability model, we require a prior specification for λ, A_i, and ϵ. Let $\lambda \sim$ inverted chi-square(ν, ξ). A choice of parameters that imposes almost no effect on the likelihood is $\nu = 0$, $\xi = 1000$; this is essentially the reference prior (Lee, 1989). Let A_i be a set of independent random variables, each with a normal distribution $\mathrm{N}(\kappa, \tau^2)$. A prior could also be placed on ϵ (Verdinelli and Wasserman, 1991), but we obtained better results with a fixed value of ϵ. The best choice of values for $(\kappa, \tau^2, \epsilon)$ depends on whether the markers under study are part of a genome scan or a candidate gene study (see below).

This model is quite convenient for making Bayesian inferences via Gibbs sampling (Carlin and Louis, 1996), with simple conditional distributions required to compute the Gibbs updates. Conditional on the data and the other parameters, the following distributions result. If $\delta_i = 1$, A_i is $\mathrm{N}(c, d)$ with $d^{-1} = (1/\tau^2 + 1/\lambda)$ and $c = \left(X_i/\lambda + \kappa/\tau^2\right) \times d^{-1}$. If instead, $\delta_i = 0$, then A_i has the prior density $\mathrm{N}(\kappa, \tau^2)$. Each δ_i is independent and is distributed as a Bernoulli with success probability

$$p_i = \frac{\phi\left\{(X_i - A_i)/\lambda^{1/2}\right\}\epsilon}{\phi\left\{(X_i - A_i)/\lambda^{1/2}\right\}\epsilon + \phi\left(X_i/\lambda^{1/2}\right)(1-\epsilon)}, \qquad (8)$$

where $\phi(\cdot)$ represents the standard normal density. λ has an inverted chi-square distribution. More precisely, $\Sigma\,(X_i - \delta_i A_i)^2 + \nu\xi/\lambda \sim \chi^2_{n+\nu}$.

Determining if δ_i is one or zero is a binary Bayesian hypothesis testing problem (Berger, 1985). When the loss is zero or one, then the rule is to choose $\delta_i = 1$ if

$$\frac{f(X_i \mid \delta_i = 1)}{f(X_i \mid \delta_i = 0)} > \frac{(1-\epsilon)}{\epsilon}, \qquad (9)$$

where $f(X_i \mid \delta_i)$ is the marginal distribution of X_i, given δ_i. The rule above is equivalent to declaring X_i an outlier whenever $(1/M)\,\Sigma_{j=1}^M p_i^{(j)} \approx (1/M)\,\Sigma_{j=1}^M \delta_i^{(j)} > 0.5$, where $j = 1, \ldots, M$ indexes the cycles of the Gibbs sampler. For $\epsilon = 0.01$ (0.0001), the test requires the likelihood of the data under the alternative to be 99 (9999) times greater than under the null model. In this sense, ϵ can be viewed as a tuning parameter—the smaller ϵ, the higher the hurdle for declaring an observation as an outlier.

Genome scan. Risch and Merikangas (1996) proposed a genome-wide association scan using SNP as a means of determining the genetic basis of complex disease. While the number of SNP required to perform an effective scan is unknown, a reasonable estimate is between 50,000 and 100,000 SNP. A frequentist approach might use Bonferroni correction to account for such a large number of hypothesis tests. In Bayesian decision theory, the multiple comparison problem must be handled via choice of prior distributions. To avoid incurring an excess of false positives for such large n, a relatively small value of ϵ is required. For a genome scan in which no prior information is available for ϵ, we suggest ϵ equal to $10/n$. In addition to adjusting ϵ, a data-dependent prior for κ helps to account for the larger number of tests performed by moving the prior for the outlier distribution to the right of the expected size of the largest order statistic obtained in n independent tests. We suggest data-dependent values for both κ and τ^2, with $\kappa = \hat{\lambda}^{1/2} \times \mathrm{E}[X_{(n)} + 1]$ and $\tau^2 = \hat{\lambda}$, where $\hat{\lambda} = \{\mathrm{median}(X_1, X_2, \ldots, X_n)/0.675\}^2$ is a robust measure of λ and $\mathrm{E}[X_{(n)}]$ is the expected value of the largest order statistic from a sample of size n from a standard normal distribution. The largest order statistic is of order $\{\log(n)\}^{1/2}$, or about $1.6\{\log(n)\}^{1/2} - 1$, in the range of interest.

Candidate genes. A slightly different approach, consisting of two stages, better utilizes the potential of a Bayesian analysis in a candidate gene study. Here we assume that the n loci consist of c biallelic polymorphisms in candidate genes and $(n - c)$ SNP dispersed throughout the genome. The SNP markers are examined for other purposes, such as a linkage study. Because the SNP markers are widely spaced, perhaps throughout the genome, we assume that they are not likely to be near enough to a susceptibility gene to exhibit a detectable level of association. Consequently, only the candidate gene markers will be tested for association.

In stage 1, the outlier test is performed. Because a small number of tests are to be performed (c versus n) and because there is prior information implicating the markers under study, a candidate gene study has the potential of yielding much more powerful inferences. The key tuning parameter is ϵ. We suggest using $\epsilon = 0.10$ and/or 0.05 for a preliminary screening test. According to Jeffrey's criterion (Kass and Raftery, 1995), loci that are declared outliers with $\epsilon = 0.05$ provide strong evidence of association and those that are outliers with the less stringent $\epsilon = 0.10$ provide substantial evidence of association. The remaining tuning parameters (κ, τ^2) are of less importance, and we suggest setting them at $(4\hat{\lambda}^{1/2}, \hat{\lambda})$. With these choices, the test should have acceptable Type I error rates.

Those candidate genes with the highest posterior probability of association and strongest associations with the disorder are the most promising ones to pursue further. At stage 2, these quantities are computed for each candidate gene determined to be an outlier in stage 1. To complete the computations, we require a subjective declaration of ϵ, which is interpreted as the prior probability a given candidate gene is associated with the disorder. For instance, if the candidate gene was strongly implicated in one or more prior studies, ϵ might be set at 0.20 or even greater. Alternatively, if the candidate gene was weakly implicated based on a single poorly designed study, $\epsilon = 0.05$ might be appropriate. For any prespecified prior probability ϵ, the posterior probability of association is computed as given by $(1/M)\Sigma_j p_i^{(j)}$. If there is cause to vary ϵ by marker, then this analysis should be performed on a locus-by-locus basis. The posterior distribution of the strength of the association, $A_i \sim \mathrm{N}(c, d)$, can be computed simultaneously.

2.2.2. *Frequentist approach.* The idea of genomic control can also be implemented without resorting to Bayesian techniques. Numerous frequentist outlier tests are applicable to this situation (see Barnett and Lewis, 1995, chapter 6). Many of these tests, however, are sensitive to swamping and masking effects. For this reason, we favor the simple, robust technique described below.

Testing. A frequentist outlier test can be derived based on the fact that $X_i = |Y_i|$ is approximately distributed as the absolute value of an $\mathrm{N}(0, \lambda)$ random variable under the null hypothesis. Although λ is unknown, with large n, it can be estimated with high precision using a robust estimator such as $\hat{\lambda}$ (defined in the previous section). When n, R, and S are large, $Y^2/\hat{\lambda}$ is approximately distributed χ^2_1 under the null hypothesis. A Bonferroni correction provides a conservative critical value for the test, $\chi^2_1(\alpha/n)$. When λ is constant across the genome, this simple adjustment will result in a test statistic with Type I error rate close to the nominal level. When λ follows a distribution across the genome with standard deviation of the same order as the mean, this adjustment will result in a test statistic with Type I error rate roughly equivalent to the nominal level.

Power. An advantage of the genomic-control methodology is that levels of heterogeneity and cryptic relatedness need not be prespecified. However, without knowledge of λ, it is difficult to design a study that attains a prespecified level of power. Because the greatest impact on the test statistic arises due to cryptic relatedness, we recommend using a fixed level of cryptic relatedness to obtain a conservative estimate of power. From (6), λ can be computed. Then N can be determined based on a test that rejects for values greater than $\lambda\chi^2_{\alpha/n}$.

3. Simulation Results

In real populations, clusters of individuals are related to varying degrees and population heterogeneity varies somewhat over loci. Even if it were practical to simulate reality, it would be difficult to summarize the simulations in a compact form. Fortunately, because F can be interpreted both in terms of the correlation due to relatedness and correlation due to population substructure, there is a simple way of generating data to evaluate the methods. As noted previously, the most extreme population heterogeneity occurs when cases and controls are sampled from distinct subpopulations. When this occurs, cases (and controls) are related to each other by a fixed degree.

To model outliers, we produce data from a multiplicative model for genotype relative risk (Risch and Merikangas, 1996) with approximate risk parameter γ (see Sasieni, 1997, Theorem 1). For risk to individuals carrying zero, one, and two alleles at a liability locus being π_0, π_1, and π_2, γ equals π_1/π_0 and γ^2 equals π_2/π_0.

By standard techniques for beta-binomials, we simulate data with the desired correlation structure: Within the population as a whole, the cases possess a fixed allelic correlation both within and across genotypes equal to F_1, but the genotypes of the cases are uncorrelated with the genotypes of the controls. Similarly, the controls are generated

Table 3

Comparison of power and Type I error rates for the genomic-control test and the standard procedure (trend tests with Bonferroni correction). For each configuration, 100 *data sets with* 400 (100,000) *loci were generated, each with* 10 *loci actually in association with the disorder [relative risk equal to γ, ϵ =* 0.01, (0.0001) *and κ =* $4\hat{\lambda}$ ($5\hat{\lambda}$)]. *The columns labeled Outliers give the average number of observations correctly declared as being in association with the disorder by the two statistical procedures; this column divided by* 10 *gives the power to detect an outlier with this level of relative risk. The columns labeled Errors give the average number of observations incorrectly declared as outliers by the two statistical procedures; this column divided by* 390 (99,990) *gives the Type I error rate per locus.*

				Genomic-control		Standard	
$R = S$	γ	F_1	F_2	Outliers	Errors	Outliers	Errors
				n = 400			
1000	1.25	0.00001	0.00001	5.45	0.39	4.44	0.14
	1.50	0.001	0.00001	9.59	0.29	10.00	3.71
	2.25	0.01	0.00001	8.51	0.40	10.00	106.21
	1.50	0.001	0.001	6.51	0.34	9.72	13.41
	2.50	0.01	0.001	9.38	0.31	10.00	114.83
	3.00	0.01	0.01	5.39	0.21	10.00	166.50
500	1.50	0.00001	0.00001	8.86	0.34	8.09	0.11
	1.75	0.001	0.00001	9.89	0.42	9.94	0.92
	2.50	0.01	0.00001	9.26	0.33	10.00	53.64
	1.75	0.001	0.001	8.84	0.40	9.83	3.76
	2.75	0.01	0.001	9.54	0.36	10.00	59.46
	3.25	0.01	0.01	6.39	0.26	10.00	107.60
100	2.00	0.00001	0.00001	5.18	0.25	3.63	0.25
	2.25	0.001	0.00001	6.62	0.32	5.99	0.12
	3.00	0.01	0.00001	6.16	0.23	9.20	3.48
	2.25	0.001	0.001	5.79	0.23	9.20	3.48
	3.50	0.01	0.001	7.80	0.24	9.81	4.16
	4.50	0.01	0.01	7.09	0.28	9.92	12.70
				n = 100,000			
1000	1.40	0.00001	0.00001	7.96	0.96	6.40	0.08
	2.75	0.01	0.00001	5.28	0.32	10.00	14,125.47
	1.75	0.001	0.001	8.12	0.68	10.00	465.12
100	3.00	0.00001	0.00001	7.48	0.72	5.84	0.04
	4.50	0.01	0.00001	5.56	0.28	9.64	36.92
	3.00	0.001	0.001	5.60	0.56	5.52	0.48

with a fixed allelic correlation equal to F_2 but unrelated to the cases. Data for cases and controls are generated independently with each population in Hardy–Weinberg equilibrium. For cases, $p = \Pr(A_1)$ is sampled from a beta distribution with parameters $\alpha = \beta = (1 - F)/(2F)$; then a binomial sample of $2R$ alleles is drawn using this value of p; these alleles are randomly paired to form genotypes. For controls, a new value of p is sampled and then a sample of $2S$ alleles is drawn using this value of p. A new value of p is generated for each locus in both the cases and controls. In each draw, the expected value of p is 1/2 except for the loci designated as outliers. For the outliers, the alleles are sampled from a binomial$\{2R, \gamma/(1 + \gamma)\}$ distribution and are randomly paired to form genotypes.

We simulate from values of F ranging from 0.01 to 0.00001. Within a randomly mating population, these values represent a range of cryptic relatedness spanning approximately second to seventh cousins. $F = 2^{-2(k+1)}$ for k-cousins. We choose fairly large values of F to account for the possibility that the cases and controls are from slightly different subpopulations.

From Table 3, it is apparent that the Type I error rate is small and quite stable for the Bayesian genomic-control test. A single false positive is obtained with probability roughly 0.33 for $n = 400$ and 0.60 for $n = 100{,}000$. Contrast this stability with the standard test (trend tests with a Bonferroni correction $\alpha = 0.10/n$), which yields wildly unstable numbers of false positives ranging from 0 to over 14,000 errors per genome scan. In general, a larger number of false positives occurs when F is larger. As expected, correlation among subjects has a strong effect on the distribution of the test statistic and this effect is most acute when the sample size is larger (i.e., the effect increases as R increases).

Another feature illustrated by the simulations is that the power of the tests decreases as n increases. This is not surprising because a good test must be more conservative to prevent a large number of false positives when a dense genome-wide scan is performed.

$F_1 = F_2 = 0.00001$ represents the ideal model because alleles are essentially uncorrelated and the test statistics are very nearly asymptotically chi-square distributed under the null hypothesis. Not surprisingly, the standard test results in very few false positives both for $n = 400$ and $n = 100{,}000$. For $n = 400$, the power of the standard test is about 10% lower than that of the genomic-control test. For $n = 100{,}000$, the loss of power is more substantial, i.e., approximately 16%.

4. Discussion

Our genomic-control methods target the detection of population-level association between marker and disease from a case–control sample. They are designed to exploit advancing technology for the detection of genes underlying human diseases, such as single nucleotide polymorphisms detected using a gene chip, a glass wafer to which is bound high-density arrays of prepooled primers for multiplex polymerase chain reaction assays. The first generation of gene chips is due in 1999 (see http://www.affymetrix.com/), and up to 100,000 SNP scattered throughout the genome are anticipated to be available within a few years (Collins et al., 1998).

For a case–control sample, population substructure and cryptic relatedness among subjects leads to overdispersion of the chi-square test statistic for association and causes spurious rejections of the null hypothesis. Under reasonable population genetic assumptions, however, this overdispersion is roughly constant across the genome, allowing for a natural correction to the case–control test statistic. Plainly, this correction comes at a cost: Case–control studies analyzed using the genomic-control approach incur a reduction in power if the sample is not independent. The larger the overdispersion parameter, the smaller is the power. Consequently, although the genomic-control method allows for the analysis of case–control samples that do not meet the independence assumption, a carefully collected sample from a homogeneous population of unrelated individuals will yield a more powerful test statistic.

In fact, the genomic-control method produces control in many ways comparable to genetic epidemiology's family-based designs (Spielman et al., 1993; Curtis, 1997). These family-based designs, which are matched case–control designs with appropriate test statistics (e.g., Laird, Blacker, and Wilcox, 1998), circumvent spurious association due to population heterogeneity. As we have demonstrated here, the genomic-control method also eliminates spurious associations due to population heterogeneity. There are other favorable features of case–control methods, which, when teamed with genomic-control methodology, make case–control a very compelling method for the genetic analysis of complex diseases. For instance, family-based designs generally are not efficient relative to case–control designs for genetic analysis of complex diseases (Risch and Teng, 1998). In addition, family-based designs require tremendous effort during the data collection phase compared with case–control studies and therefore cost far more to implement.

The proposed genomic-control method is built on a Bayesian probability model. This model easily accommodates overdispersion due to heterogeneity and relatedness. With the help of tuning parameters, the method also scales as the size of the genome scan increases, alleviating concerns over multiple testing. By adjusting the tuning parameters, the test can be scaled to have the desired Type I error rate. We provide default values that result in fairly low levels of false positives; however, if larger numbers of false positives can be tolerated, then the tuning parameters (particularly ϵ) can be adjusted to enhance the power of the test. (Software to implement genomic-control methods and select tuning parameters are available from the authors on request.)

We also described a frequentist version of the genomic-control approach. In many cases, such as our simulations, the Bayesian and frequentist methods will behave similarly. As witnessed by our suggested treatment of candidate gene analyses, however, the Bayesian approach has the advantage of being readily extended to solve more complex problems.

From the simulation study, it is clear that the genomic-control method performs substantially better than the standard method for a wide spectrum of conditions. When the sample is approximately independent, the genomic-control has greater power than that obtained by the standard procedure, but it also has a slightly greater number of false positives due to the choice of tuning parameters. Of much greater importance is the comparison of the procedures when the samples are not independent. Here we find that the genomic-control approach maintains a nearly constant low level of false positives, while the standard procedure has a wildly unpredictable level of errors. The Type I error rate for the standard procedure is especially large when the sample size is large because of the cumulative effect of the violations of independence in the sample. The power of both methods naturally declines as n, the number of markers tested, increases. This is a predictable result of the need to control for a greater risk of false positives.

We have deferred to this point a discussion of the impact of the heterogeneity of F, due to population substructure, on the genomic-control method. Clearly, F does vary in many settings, and the degree to which it varies depends on the design of the case–control study. Intuitively, the effect of this variation is to increase the variance of the test statistics, thereby increasing the value of λ. The net effect on the genomic-control procedure is to decrease its power. From some simulation analyses, it appears that λ is overestimated, and therefore the genomic-control method maintains a small false positive rate even in the presence of variation in F. Thus, because of its impact on the power of the test, it is important to design case–control studies to limit the size of F (and, implicitly, the variance of F).

Currently, the genomic-control approach is limited to biallelic markers for three reasons. First, a stronger case can be made for nearly constant overdispersion in this setting. Without this, the approach loses much of its power and appeal. Second, with only two alleles, it is not necessary to specify which allele is potentially associated with the disorder. Third, the approach is based on the comparison of test statistics across the genome to find outliers. This comparison requires that the tests all follow the same distribution under the null hypothesis. With differing allele counts, the test statistics would have differing degrees of freedom. This simple version of the genomic-control approach ignores potential spatial dependence in the test statistics. A more powerful approach could be designed that incorporates the spatial configuration. Such an approach is one focus of our current research.

Acknowledgements

The authors are grateful to the University of Pittsburgh Human Genetics Journal Club for helpful discussions. This research is supported by NIH grants NIMH 57881, NIMH 56193, and NSF grant DMS 9803433.

Résumé

Un panel dense de polymorphismes bialléliques (SNP) couvrant le génome et une méthode efficace pour tester les génotypes SNP sont attendues dans un futur proche. Une question primordiale est comment utiliser efficacement ces techniques pour identifier les gènes affectant la susceptibilité à des désordres complexes. Pour arriver à cet objectif, nous proposons une méthode statistique qui a plusieurs propriétés optimales: elle peut être utilisée avec des données cas-témoin ou encore, comme dans les études familiales, des contrôles pour l'hétérogénéité de la population; elle est insensible aux violations habituelles aux hypothèses des modèles, comme les observations n'étant pas strictement indépendantes; et, en utilisant des méthodes bayesiennes de détection de points éloignés, elle évite la nécessité d'utiliser une méthode de correction de Bonferroni pour tests multiples, aboutissant à de meilleures performances dans beaucoup de situations tout en contrôlant le risque de faux positifs. Les performances de notre méthode de "contrôle génomique" est plutôt satisfaisante pour des effets plausibles de gènes de susceptibilité, ce qui est de bon présage pour les futures analyses génétiques de désordres complexes.

References

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11,** 375–386.

Barnett, V. and Lewis, T. (1978). *Outliers in Statistical Data.* New York: Wiley.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis,* 2nd edition. New York: Springer-Verlag.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* New York: Chapman and Hall.

Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994). *The History and Geography of Human Genes.* Princeton, NJ: Princeton University Press.

Chakraborty, R. and Jin, L. (1992). Heterozygote deficiency, population substructure and their implications for DNA fingerprinting. *Human Genetics* **88,** 267–272.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. (1998). New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282,** 682–689.

Crow, J. F. and Kimura, M. (1970). *An Introduction to Population Genetics Theory.* New York: Harper and Row.

Curtis, D. (1997). Use of siblings as controls in case–control studies. *Annals of Human Genetics* **61,** 319–333.

Devlin, B., Risch, N., and Roeder, K. (1993b). Statistical evaluation of DNA fingerprinting: A critique of the NRC's report. *Science* **259,** 748–749, 837.

Elandt-Johnson, R. C. (1971). *Probability Models and Statistical Methods in Genetics.* New York: John Wiley.

Falk, C. T. and Rubinstein, P. (1987). Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics* **51,** 227–233.

Kass, R. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90,** 773–795.

Laird, N. M., Blacker, D., and Wilcox, M. (1998). The sib transmission/disequilibrium test is a Mantel–Haenszel test. *American Journal of Human Genetics* **63,** 1915.

Lee, P. M. (1989). *Bayesian Statistics: An Introduction.* London: Edward Arnold.

Lewontin, R. C. and Krakauer, J. (1973). Distribution of gene frequencies as a test of the theory of selective neutrality of polymorphisms. *Genetics* **74,** 175–195.

Li, C. C. (1972). Population subdivision with respect to multiple alleles. *Annals of Human Genetics* **33,** 23–29.

Morton, N. E. and Collins, A. (1998). Tests and estimates of allelic association in complex inheritance. *Proceedings of the National Academy of Science, USA* **95,** 11389–11393.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **255,** 1516–1517.

Risch, N. and Teng, J. (1998). The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. *Genome Research* **8,** 1273–1288.

Robertson, A. (1975). Gene frequency distribution as a test of selective neutrality. *Genetics* **81,** 775–785.

Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size. *Biometrics* **53,** 1253–1261.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52,** 506–516.

Verdinelli, I. and Wasserman, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler. *Statistics and Computing* **1,** 105–117.

Wang, D. G., Fan, J. B., Siao, C. J., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077–1082.

Received February 1999. *Revised June* 1999.
Accepted July 1999.