Unbiased Methods for Population-based Association Studies

B. Devlin\textsuperscript{1}, Kathryn Roeder\textsuperscript{2}, and Silviu-Alin Bacanu\textsuperscript{1}

\textsuperscript{1} Department of Psychiatry  
University of Pittsburgh

\textsuperscript{2} Department of Statistics  
Carnegie Mellon University

Address for correspondence and reprints:

Bernie Devlin, Department of Psychiatry, University of Pittsburgh School of Medicine, 3811 O’Hara Street, Pittsburgh, PA 15213. E-mail: devlinbj@msx.upmc.edu

Kathryn Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: roeder@stat.cmu.edu

Running Title: Population-based Association Studies

Send proofs to: Bernie Devlin, Department of Psychiatry, University of Pittsburgh School of Medicine, 3811 O’Hara Street, Pittsburgh, PA 15213
Abstract

Large, population-based samples and large-scale genotyping are being used to evaluate disease/gene associations. A substantial drawback to such samples is the fact that population substructure can induce spurious associations between genes and disease. We review two methods, called genomic control (GC) and structured association (SA), that obviate many of the concerns about population substructure by using the features of the genomes present in the sample to correct for stratification. The GC approach exploits the fact that population substructure generates ‘overdispersion’ of statistics used to assess association. By testing multiple polymorphisms throughout the genome, only some of which are pertinent to the disease of interest, the degree of overdispersion generated by population substructure can be estimated and taken into account. The SA approach assumes that the sampled population, while heterogeneous, is composed of subpopulations that are themselves homogeneous. By using multiple polymorphisms throughout the genome, this ‘latent class method’ estimates the probability sampled individuals derive from each of these latent subpopulations.

GC has the advantages of robustness, simplicity, and wide applicability, even to experimental designs such as DNA pooling. SA is a bit more complicated, but has the advantage of greater power in some realistic settings, such as admixed populations or when association varies widely across subpopulations. It, too, is widely applicable. Both also have weaknesses, as elaborated in our review.

Key Words: case-control study, Genomic Control, latent class model, linkage disequilibrium, population substructure, Structured Association


1 Introduction

The challenge for genetic epidemiologists during the twenty-first century is to disentangle the genetic and environmental basis of complex disease. We well recognize the difficulty. These diseases are complex because they have both genetic and environmental bases, and the individual components of these bases can be quite subtle (Risch 2000). For this reason, many early studies of complex disease, especially those relying strictly on the analysis of genetic linkage, were probably under-powered.

In the face of such complexity, we must employ efficient study designs. For general populations, ideal designs are those amenable to recruitment of large samples, thereby increasing power, and amenable to seamless modeling of both environmental and genetic factors, thereby capturing as many of the sources of variability as possible.

One solution is provided by association studies (Risch and Merikangas 1996). Human populations, however, often exhibit substructure due to non-random mating, such as that based on proximity or culture. This substructure, or stratification, creates concern about the validity of association tests. Association analyses can be conducted by using either family-based samples or population-based samples. By using the transmission of alleles from parents to offspring, the ‘family-based’ approach completely obviates concerns about population substructure. Since Spielman et al.’s (1993) seminal paper, and Ewens and Spielman’s (1995) more rigorous demonstration of its robustness to substructure, the number of family-based tests has grown tremendously. Yet, by definition, family-based association studies require families, and this feature has advantages and drawbacks. A substantial advantage is the robustness to population structure; a substantial disadvantage is that the design mitigates against the recruitment of large samples, which in turn limits power (Risch and Teng 1998). Even for the same sample sizes, family-based studies are less powerful than population-based studies for some – but not all – settings (Ewens and Spielman 1995; Bacanu et al., 2000). One could argue that what is required is some means of exploiting the facile and powerful case-control designs while simultaneously providing the protection against false positives, due to population substructure, provided by family-based designs.

Such an alternative solution springs from recent innovations in technology. With the promise of relatively-inexpensive genotyping, it is possible to account for the impact of substructure by using the distribution of markers in the sampled genomes – what we call “Genomic Control” or GC (Devlin and Roeder 1999). For example, for a case-control anal-
ysis of candidate genes, one GC approach computes $\chi^2$ test statistics for independence for both null and candidate loci. By using the variability and magnitude of the test statistics observed at the null loci, which are inflated if population stratification or unmeasured genetic relationships among the subjects (cryptic relatedness) exist, a multiplier is derived to adjust the critical value for significance tests for candidate loci (Devlin and Roeder 1999; Bacanu et al., 2000). In this way, GC permits analysis of stratified case-control data without an increased rate of false positives. If population stratification and cryptic relatedness were absent, then GC approximates a standard test of independence for a case-control design.

A related approach called “Structured Association” (SA) has been developed by Pritchard et al. (2000a), who proposed using marker loci unlinked to the candidate genes under study to infer subpopulation membership. They propose using latent class analysis to control for the effect of population substructure. In essence, using null loci, the number of subpopulations $K$ is estimated and then each subject’s probability of membership in each of $K$ subpopulations is estimated (Pritchard et al. 2000b). The key assumption of the latent class model is that, conditional on subpopulation, there is no confounding due to population substructure.

In this review, we outline the basic theory behind GC and SA, describe the various implementations to date, compare the different methods, and then discuss some open questions regarding these methodologies.

2 Confounding in Population-based Samples

It is well known that samples from a structured population can produce association between unlinked loci (Ewens and Spielman 1995; Knowler et al. 1998), which we call confounding. Confounding can have two effects on test statistics, they can be biased and/or over-dispersed, and both cause an increase in false positives over the nominal level. While bias can be a critical factor for traditional epidemiological studies, over-dispersion appears to be the dominant consequence of confounding in genetic studies (Devlin et al., in press).

In this section, we explore the impact of confounding due to population substructure via a case-control study to assess association between alleles at disease and marker loci, both of which are biallelic. To do so we define several quantities: let $c$ be an unobservable indicator of subpopulation membership, $c = 1, \ldots, K$; $D$ be a binary indicator of disease status; $A$ and $a$ be the allele labels at the marker loci; and $F_{st}$ denote the correlation among alleles drawn from a common subpopulation (Wright 1969). $F_{st}$ is greater than zero when the alleles vary
in their frequency across subpopulations. It is specifically defined as the variance of the allele frequencies standardized by the Bernoulli variability of the pooled allele frequency, \( \hat{p}(1 - \hat{p}) \).

In a case-control setting the response is genotype and the covariate is disease status, \( D \). We focus on a test statistic based on an additive genetic model and define \( \delta \) to be the case-control effect of \( A \), \( \delta = P(A|D = 1) - P(A|D = 0) \). Under the null hypothesis \( \delta \) is non-zero solely due to bias. For simplicity of exposition, assume that an equal number \( N \) of case and control subjects have been sampled. Let \( X_i \) denote the number of \( A \) alleles in the \( i \)'th case subject and \( Y_j \) denote the same for the \( j \)'th control subject. \( \hat{\delta} = (\sum_i X_i - \sum_j Y_j)/(2N) \) estimates the case-control effect. For an independent identically distributed (i.i.d.) sample with \( p = P(A) \), \( \text{Var}[\hat{\delta}] = \sigma^2 = \hat{p}(1 - \hat{p})/N \). This variance is estimated by \( \hat{\sigma}^2 = \hat{p}(1 - \hat{p})/N \).

Now to model the effect of population substructure on \( \delta \), assume \( p_c = P(A|C = c) \) is an i.i.d. random variable, \( c = 1, \ldots, K \), with mean \( p \) and variance \( F_{st}p(1 - p) \). It follows that \( E[\delta] = 0 \); i.e., averaged over randomly selected subpopulations, the case-control bias has mean zero. The variance of the bias is \( E[\delta^2] = \sigma^2 \times \eta^2 \) (Devlin et al. \textit{in press}). This quantity is zero if there is no variability in allele frequencies due to population substructure; no variability in prevalence of the disease across subpopulations; or no correlation in the variability of marker allele frequencies with disease prevalence across subpopulations. Clearly the first two factors are unlikely, but the latter factor, low correlation, often holds when the number of subpopulations is large (Chakraborty and Jin 1992; Wacholder et al. 2000), and this is the typical case for human populations.

In a substructured population the Wahlund effect predicts the covariance between alleles within a subject equals \( F_{st}p(1 - p) \), which inflates the variance of the allele counts within an individual. For a substructured population with no inbreeding \( F_{st} \) is also the correlation between alleles from members of the same subpopulation. As a consequence of this correlation, the usual statistical test for association can result in a rate of false positives exceeding the nominal level. To see this, let \( \zeta_c = P(C = c|D = 1) \) and \( \omega_c = P(C = c|D = 0) \) denote the fraction of cases and controls from each of the \( c = 1, \ldots, K \) subpopulations. It follows from Roeder and Devlin (1999) that

\[
\text{Var}[\hat{\delta}] = \sigma^2 \times \left\{ 1 + NF_{st} \sum_c (\zeta_c - \omega_c)^2 - 2F_{st} \right\} = \sigma^2 \times \tau^2, \tag{1}
\]

where \( \tau^2 \) is the inflation of variance due to correlated alleles. Because \( \tau^2 \) increases as a function of \( N \) it can be sizable even for small allelic correlations. Because cases or controls may be related, albeit cryptically, (1) can be generalized to incorporate cryptic relatedness.
(Devlin and Roeder 1999).

Consider computing \( \chi^2 \) tests for association at a number of independent markers, indexed by \( k \), across the genome. Note that \( S_k \equiv (\hat{\delta}_k / \delta_k)^2 \) is equivalent to the desired \( \chi^2 \) test statistic if the additive model is used. Based on the results concerning the bias and variance of \( \hat{\delta}_k \) obtained thus far, it can be shown that \( S_k \) is approximately distributed as a scaled \( \chi^2 \), \((\eta^2 + \tau^2)\chi^2_1\), under reasonable assumptions (Devlin et al. in press). We conclude that the effect of confounding is to inflate the \( \chi^2 \) test statistic by a factor \( \lambda = \eta^2 + \tau^2 \), which is typically greater than 1 if population substructure, admixture or cryptic relatedness are present in the sample. Thus confounding can lead to bias and over-dispersion, both of which can produce excess false positives when testing for association, especially when \( N \) is large.

3 Genomic Control and Structured Association

Based on the observation that population substructure induces dependence of alleles at a locus, many population-based association studies also tested the samples for violations of Hardy-Weinberg equilibrium. Similarly Pritchard and Rosenberg (1999) suggested evaluating a large number of loci unlinked to the candidate gene of interest to determine if there is evidence of association indicating substructure. These approaches have some drawbacks: (1) because all human populations are substructured to some degree, association will be detected, almost surely, as the sample size or the number of loci tested increases; and, (2), when association is detected, how does one proceed with the study?

In 1999, we proposed an alternative approach – Genomic control or GC. The general principle of GC is to use individual genomes, as presented in the sample, to account for the confounding due to substructure and more-direct relatedness. Building on results from evolutionary theory [e.g. Wright (1969), Lewontin and Krakauer (1973)], Devlin and Roeder (1999) demonstrated that the effects of cryptic relatedness and population substructure on test statistics of interest are essentially constant across the genome, under certain conditions. We suggested using “null” markers (e.g., polymorphisms unlikely to affect liability) across the genome to estimate the effect of confounding and then removing the effect from the association test statistic.

Since the GC method was introduced, the Structured Association or SA method has been developed. Pritchard et al., (2000a) proposed using loci unlinked to the candidate genes under study to infer subpopulation membership. The idea is that, conditional on
subpopulation, there is neither bias nor excess variance due to population substructure. These authors construct a two-stage procedure: in the first stage each subject’s probability of membership in each subpopulation is estimated (using methods in Pritchard et al. 2000b); in the next stage, a test of association is conducted within subpopulations. Pritchard et al.’s (2000a) method falls under the rubric of latent class models (the subpopulations are the unknown, latent classes). For related work, see Satten et al., (2001), who bypass the two-stage approach by simultaneous estimation of latent classes and association, and Schork et al. (2001), who applied this general idea to 44 microsatellite loci used in a renal failure study.

3.1 The GC Approach

We initially illustrate the over-dispersion model for a biallelic marker and a case-control sample, and then outline how these results extend to some other experimental designs.

GC for Case-control Studies with Biallelic Markers. Assume a set of biallelic loci are evaluated. We can test for association at any particular locus using a $\chi^2$ test based upon the $2 \times 2$ allelic table. This is the statistic, $S_k$, discussed in §2. Under the null hypothesis, for large $N$, $S_k$ is approximately distributed as a scaled, $\chi^2$ random variable with scaling parameter $\lambda = \eta^2 + \tau^2$. By evaluating multiple loci, $\lambda$ can be estimated and tests for association can be adjusted by dividing the test statistic by $\hat{\lambda}$ (see Devlin and Roeder 1999; Bacanu et al. 2000).

To estimate $\lambda$ two choices are natural: a robust estimator such as the median of the $\chi^2$ test statistics, divided by 0.456 (Devlin and Roeder 1999), or the mean (Reich and Goldstein 2001). Because there is sampling variability in $\hat{\lambda}$, it is natural to bound the correction factor using $\max(\hat{\lambda}, 1)$ as in Bacanu et al. (2000). Reich and Goldstein (2001) recommend a more conservative correction to account for the sampling variability in $\hat{\lambda}$; in practice this leads to estimates for $\lambda$ that are substantially larger than the truth, on average. In simulations, Bacanu et al. (2000) found that the bounded median estimator performed well when 50 or more null loci were utilized, and that it was conservative, on average, when only 20 null loci were available.

GC for Quantitative Trait Studies. Let $Y$ be a quantitative outcome variable that is influenced by the genotypes at numerous loci. In general we can simultaneously test the effect of multiple loci (Bacanu et al., 2001), but here we discuss only the simplest case. For a single locus consider the model $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$, where $X_{i1}$ is the number of $A$ alleles in the $i$’th individual, and test whether the slope is different from zero. The usual estimator
of the parameter of interest is the least squares estimator, \( \hat{\beta}_1 \). Define \( SE_{ind}[\hat{\beta}_1] \) as the usual standard error term that would be obtained assuming that the \( Y_i \)'s are independent, i.e., the term that would be obtained directly from any statistical regression package. Due to confounding arising from substructure, Bacanu et al. show that the actual variance of \( \hat{\beta}_1 \) is inflated by both bias, which is usually small, and correlation among the subjects in the study. The inflation of the test statistic \( QT = [\hat{\beta}_1/SE_{ind}(\hat{\beta})]^2 \) is identical to that for the case-control setting. Thus \( QT \) is approximately distributed \((\eta^2 + \tau^2)\chi^2_1 = \lambda\chi^2_1\) and the problem is amenable to the GC approach.

**GC for Case-control Studies with Multi-allelic Markers and Haplotypes.** Studies testing association of disease with multi-allelic markers or haplotypes can be more informative than tests using biallelic markers. The omnibus \( \chi^2 \) test for association is the usual choice for such studies, yet, if this test has many degrees of freedom, it can have small power even for substantial associations. To combat this problem we consider one measure of association that uses a single degree of freedom among the many possible statistics in this class.

Consider a certain segment of a chromosome: two haplotypes of that particular segment are defined as matching if all their alleles are the same. For \( L_k \) haplotypes in chromosomal segment \( k \), with corresponding haplotype frequencies as \((\pi_{a1(k)}, \pi_{a2(k)}, \ldots, \pi_{aL_a(k)})\) in cases and \((\pi_{u1(k)}, \pi_{u2(k)}, \ldots, \pi_{uL_u(k)})\) in controls, the probability of two case haplotype segments matching is \( Pr(\text{Hap}_i \text{ and Hap}_j \text{ match}) = \sum_{l=1}^{L_k} \pi_{al(k)}^2 \) and likewise for controls. A suitable test statistic could be based upon the difference in the matching probability, \( T_k = \sum_{l=1}^{L_k} \pi_{al(k)}^2 - \sum_{l=1}^{L_k} \pi_{ul(k)}^2 \), which is sensitive to detecting excess matching in case subjects versus control subjects. For this reason we expect this statistic to perform well even when susceptibility alleles are common. For sufficiently large samples, \( T_k \) follows a normal distribution. By extending the reasoning applied for biallelic markers to this setting it is possible to show that the GC procedure extends to this problem (Tzeng et al., 2001).

**GC with Pooled DNA.** The GC approach can be applied directly to “pooled DNA”, wherein DNA from many cases and many controls are mixed before genotyping and only the allele frequencies for cases versus controls can be determined. Pooled DNA amplification can facilitate efficient, cost-effective, high-resolution genome screening for detection of disease loci by association (Barcellos et al. 1997; Teng and Risch 1998). We discuss this extension in the context of case-control studies with biallelic marker loci.
In its simplest implementation estimates of the allele frequencies are obtained from the pooled experiment and then estimates of $\hat{\delta}_k$ and $\hat{\sigma}_k$ are computed. An estimate of $S_k$, which we call $\tilde{S}_k$, follows. Directly applying the GC method to $\tilde{S}_1, \tilde{S}_2, \ldots, \tilde{S}_n$ is valid under either of two conditions: (i) the variance in the allele frequency estimates due to pooling is small relative to the sampling error; or (ii) the variance due to pooling is proportional to $p_k(1-p_k)$.

A common design for estimating pooled allele frequencies uses multiple pools of the same individuals DNAs (e.g., Buetow et al. 2001). By averaging over the separate trials, the investigator obtains both better estimates of the allele frequencies and estimates of the variability across pools. In terms of GC, then, a simple algorithm unfolds: for each locus $k$, collect the the pooled measurements repeatedly; utilizing the repetition, estimate the variance due to pooling, call it $\gamma_k^2$; finally define a modified test statistic $S_k^* = \bar{\delta}_k^2 / (\hat{\sigma}_k^2 + \gamma_k^2)$ and apply the GC approach to this test statistic.

**Detecting Outliers.** For all the methods described so far, we require an estimate of one or two parameters, such as $\lambda$ or $\mu_0$ and $\tau^2$, which we assume are constant across the genome for all null loci. Association studies are performed for both candidate genes and genome scan studies. Because the bulk of the measured loci will naturally be null, provided a robust estimator is chosen, the parameters can be estimated using all of the data. For example, when the test statistics are distributed as $\lambda \chi^2_1(0)$ a robust estimator of $\lambda$ is $\hat{\lambda} = \text{median}(S_1, S_2, \ldots, S_n) / 0.456$. Devlin and Roeder (1999) present a Bayesian procedure for performing GC for a genome scan that is more powerful, but also more complicated to implement. Tzeng et al., (2001) presents a non-Bayesian method for outlier detection based upon the concept of limiting the false discovery rate (Benjamini and Hochberg 1995) rather than limiting the Type I error rate; that is, controlling the expected fraction of rejections that are false, rather than controlling the probability of a false rejection. In exchange for redefining the criterion for significance, considerable power can be gained in some circumstances, while the fraction of false positives is still controlled.

### 3.2 The SA Approach

Rather than correcting for population substructure by modeling the over-dispersion parameter, the general idea behind latent class models is to cluster the samples into groups with similar genetic ancestry and then to condition upon this ancestry covariate. For Pritchard et al.’s (2000b) method, a Bayesian clustering program is run to determine both the number $K$ of subpopulations within the population and the membership probability vector for
each individual. Model choice for $K$ is performed by running a Markov chain separately at different values of $K$; an approximate method is used to estimate posterior probabilities for each value of $K$. The program STRUCTURE performs this task (Pritchard et al. 2000b). As output one obtains a vector $q = (q_1, \ldots, q_K)$ that indicates an estimate of the fraction of the individual’s genome that originated from each of the $K$ subpopulations. This provides a flexible way of capturing various patterns of admixture.

Given an estimated membership vector $q_i$ for each of the subjects in the study, the next step is to compute a likelihood ratio test based upon computing the likelihood of the data under the null and alternative hypotheses (Pritchard et al. 2000a). Under the null, it is assumed that the genotype of the candidate gene is independent of the case/control status of the subject. Under the alternative, the model is quite flexible, allowing any genotype to be associated with case status; the model does not restrict the association to be constant across subpopulations. In this second phase of the analysis both $K$ and the membership vectors are treated as known quantities. To assess significance a smoothed bootstrap is recommended. The program STRAT uses the output from STRUCTURE as input in conducting a test for association.

Satten et al. (2001) modify the approach of Pritchard et al. (2000a,b) by developing a method that simultaneously estimates subpopulation membership and the effect of the liability mutation. Although their modeling assumptions are quite similar to Pritchard et al. (2000a,b), their implementation differs substantially. For instance, they use an EM algorithm and Swartz’s criterion to determine group membership and $K$. These authors base their inferences upon a complex latent class model with a constant odds ratio representing the effect of the candidate gene across subpopulations.

There are subtle statistical advantages to a simultaneous implementation. First, with a two-stage implementation, a bias is introduced into the estimates of group membership because it implicitly assumes, a priori, that the probability of group membership is constant across the two types of subjects (cases and controls). Second, the variance in the estimation of group membership and the number of groups is ignored when a smoothed bootstrap procedure is utilized with the two-stage process. This could lead to an inflated Type I error rate; however, these biases are often negligible. The improvements sought by Satten et al. come at a price of restrictions on the type of hypotheses that can be tested. Moreover, at the present time there is no software to implement the method.
4 Comparisons

Although similar in conception, GC and SA have different strengths and weaknesses. The SA approach is more ambitious, attempting to infer the genetic ancestry of each individual sampled, and consequently requires more null loci to ensure a successful outcome (Pritchard and Donnelly, in press). When too few null loci are used, $K$ may be underestimated, which can cause the method to be anti-conservative. Assuming the population structure can be reconstructed, however, this approach has the capacity to model different effects in different subpopulations. This extra flexibility leads either to a gain in power or a loss, depending upon whether it is needed to model the data appropriately. Certainly it is biologically plausible that effects could vary across subpopulations. It is difficult to predict this a priori.

In a sense, the GC approach is more general because it corrects for confounding due to cryptic relatedness as well as population substructure and admixture. The SA approach assumes that the observations are independent, conditional upon the inferred genetic associations due to population substructure and admixture. Consequently the SA approach is not valid for isolated or inbred populations which tend to have substantial amounts of cryptic relatedness.

A limitation of the GC approach is that it assumes the effect of population structure and admixture is approximately constant over all loci. Although this assumption is generally supported both theoretically and empirically (e.g. Bacanu et al. 2000), it is inappropriate if the loci under study are under strong selective pressure (Robertson 1975).

Here we consider the performance of GC as described in Bacanu et al. (2000) and SA as implemented in STRAT (Pritchard et al. 2000a) for case-control studies with bi-allelic markers. A number of simulations studies have been conducted to explore the Type I error rates of the GC and SA methods (Bacanu et al. 2000, Pritchard et al. 2001a), and to compare the power of the various methods: GC with TDT (Bacanu et al. 2000), SA with TDT (Pritchard et al. 2001a) and SA with GC (Pritchard and Donnelly, in press).

Simulations can be conducted under a variety of circumstances depending upon the level of substructure desired. A challenge to conducting an extensive simulation study of SA is the amount of computational effort required to obtain an estimate of $K$ using the program STRUCTURE. For this reason, in simulations $K$ is generally determined from a modest number of datasets, say 20, and then taken as given for the remainder of the investigation. Pritchard et al. (2001a) generated data with 2 subpopulations and a considerable amount of
differentiation using the coalescent model. Under these circumstances the SA approach performed quite well and achieved the appropriate Type I error rates. Using a data-generating model similar to that implemented by Bacanu et al. (2000), Pritchard and Donnelly (in press) investigate 3 subpopulations, $N = 200$ and differentiation $F_{st}$ ranging from 0.01 to 0.04. They find that SA performance depends upon the number of null loci utilized. With 200 or 1000 null loci, $K$ is correctly estimated to be 3 and the test achieve the appropriate Type I error rate. However, with only 50 loci, in 20% of the trials $K$ was estimated to be 1 and the rejection rates were too large. For the remaining 80% of the trials, $K$ was estimated to be 2, but the average rejection rate was equal to the nominal rate.

True populations usually consist of many different subpopulations, often differing in subtle ways from each others. In addition the effect of population substructure increases with the sample size. To investigate the performance of STRAT for large sample sizes ($N = 2000$) and many subpopulations ($K = 12$) we simulated 10 datasets using methods similar to those described in Bacanu et al. (2000). Each population consisted of 12 subpopulations of equal frequency and, of these 12, 6 came from one major ethnic group and the remaining 6 came from another major ethnic group. The $F_{st}$ between populations was set at 0.03 and the $F_{st}$ within populations was set at 0.01. Using only 40 null loci, we found that STRUCTURE estimated $K = 2$ for each population. Extrapolating that $K = 2$ for each additional data set, we analyzed at total of 170 datasets. Overall we found that STRAT produced the nominal level of rejections, approximately, suggesting that a finite approximation to the near continuous level of substructure present in realistic populations can protect against confounding. These results differ somewhat from those obtained by Pritchard and Donnelly (in press). We attribute some of the differences to the larger sample sizes used in our simulations (2000 vs. 200) and, possibly, the limited number of repetitions performed in our simulations to determine $K$.

In all of the published simulation studies the average rejection rates for GC were close to the nominal values (Bacanu et al. 2000, Pritchard and Donnelly, in press, Bacanu et al. 2001). We have simulated innumerable conditions with both extreme and modest levels of structure and found no setting where the method is anti-conservative on average (data not shown).

Regarding power, the best choice of analysis for any particular dataset is unclear. With a moderate amount of substructure both methods appear to be superior to the TDT in that the power, per person recruited, is generally greater for both the GC and SA methods than it
is for the TDT (Bacanu et al. 2000; Pritchard et al. 2000a; Pritchard and Donnelly in press). The notable exception occurs in admixed populations where the TDT receives an extra boost in power (Ewens and Spielman 1995). Limited simulations have been performed by Pritchard and Donnelly (in press) comparing SA and GC. From these simulations it appears that GC is more powerful than SA when the effect is nearly constant across subpopulations; however, if the effect varies strongly across subpopulations, SA is more powerful than GC.

5 Concluding Remarks

Although there are some examples of confounding due to population substructure leading to spurious association (Knowler et al. 1998), there are fewer documented cases than one would expect, given the level of concern voiced in the literature. Frequently, reported associations have not been replicated, but it is not clear if this is due to a lack of power (Morton and Collins 1998), failure to correct for multiple testing (Teng and Risch 1998), publication bias, population substructure, or even differing environmental conditions or epistatic effects. Regardless of differences between the actual and perceived impact of population substructure on studies of association, statistical methods that control for this effect will be a useful addition to the arsenal of tools for gene discovery, especially if they do so without having a substantial impact on the power of the test when corrections are unwarranted.

Genomic Control (GC) and Structured Association (SA) are new methods for robust inference of association between alleles at a disease and marker locus. When samples are drawn from heterogeneous human populations, spurious associations between alleles at unlinked loci are generated. GC and SA attempt to eliminate these spurious associations. SA models the population substructure via latent variables that code for subpopulation membership; this is the approach taken by Schork et al. (1999, 2001), Pritchard et al. (2000a), Satten et al. (2001), Ripatti et al. (in press) and Sillanpaa et al. (in press). GC models the effect of confounding as a nuisance parameter in a simple over-dispersion model. This approach is taken by Devlin and Roeder (1999), Bacanu et al. (2000, 2001), Devlin et al. (2000, in press), Reich and Goldstein (2001) and Tzeng et al. (2001). GC is startlingly simple to implement; with the help of publically available software (www.stats.ox.ac.uk/mathgen/software.html) the SA method is also easy to implement.

In this way, we hope these new methods will facilitate the search for disease alleles affecting liability to human diseases that have complex genetic and environmental bases. In this
review we have described how GC can be applied to case-control studies using biallelic and multi-allelic loci, haplotypes, quantitative traits and pooled DNA (§4.1). We also described the SA concept (§4.2) which has been applied to case-control studies only, thus far, but clearly can be extended in many directions.

GC, as we have developed it, is not without caveats. As we analyze extensively herein (§2), we assume the impact of substructure, both in terms of variance and bias, is constant across the genome. Devlin and Roeder (1999) show that some variability of $F_{st}$ would not be problematic, but substantial variability would seriously compromise the power of any study if it were not accounted for *a priori* in the statistical model. For example, amalgamating samples from populations with distinct recent histories, such as sub-Saharan Africans, Europeans, and Native Americans, without accounting for those histories in the statistical model would be foolhardy. For more subtle situations, Chakraborty and colleagues (R. Chakraborty, personal communication) have extended GC to the case of variable $F_{st}$.

Another thorny issue for GC is strong selection. If loci were under strong, subpopulation-specific selection, and these were the targeted loci for the association analysis (as opposed to the null loci used to control for population substructure), then GC would fail to exert adequate control. Alternatively, the SA approaches (Pritchard *et al.*, 2000; Satten *et al.*, 2001) to genomic control would be robust to some forms of selection if subpopulation membership can be adequately reconstructed. While such selection cannot be ruled out, we view strong selection to be unlikely for most loci that are candidates for complex human disease.

SA also has caveats. Two conditions must be met: the primary source of confounding must be due to population substructure and/or admixture, and the marker data must be sufficiently powerful to infer the number of subpopulations within the population. From simulation studies it is clear that ignoring minor structure has no noteworthy consequences, but failing to detect subpopulations that are substantially differentiated can lead to unacceptable rates of false positives.

Despite these caveats, these two new approaches offer great promise for the study of complex disease. In most cases of adult-onset disease, both approaches facilitate recruitment relative to family-based analysis, and large numbers of participants are essential to find loci of minor effect on phenotypes. Moreover, both offer a test that is more powerful than family-based tests under many circumstances.

Acknowledgements
This research was supported by National Institute of Health grants MH57881 and National Science Foundation grant DMS-9803433.
References


mapping for quantitative traits in a mixture of two populations. Genet Epidemiol (in 
press).

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: 
the insulin gene region and insulin-dependent Diabetes Mellitus (IDDM). Am J Hum 
Genet 52:506-516.

Mellon University.

studies of common genetic variants and cancer: quantification of bias. J Natl Cancer 
Inst 92:1151-1158.