Genomic control, a new approach to genetic-based association studies

B. Devlin[1], Kathryn Roeder[2], and Larry Wasserman[2]

[1] Department of Psychiatry
University of Pittsburgh

[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: Genomic Control

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

Key Words: population substructure, bias, case-control study, overdispersion, latent class model

Abstract

During the past decade, mutations affecting liability to human disease have been discovered at a phenomenal rate, and that rate is increasing. For the most part, however, those diseases have a relatively simple genetic basis. For diseases with a complex genetic and environmental basis, new approaches are needed to pave the way for more rapid discovery of genes affecting liability. One such approach exploits large, population-based samples and large-scale genotyping to evaluate disease/gene associations. A substantial drawback to such samples is the fact that population heterogeneity can induce spurious associations between genes and disease. We describe a method called genomic control (GC), which obviates many of the concerns about population substructure by using the features of the genomes present in the sample to correct for stratification. Two general approaches to population-based association studies are now available. 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 other approach, called Structured Association (SA), assumes that the sampled population, while heterogeneous, is composed of subpopulations that are themselves homogeneous. By using multiple polymorphisms throughout the genome, the SA method probabilistically assigns sampled individuals to these latent subpopulations. We review in detail GC. In addition to outlining the published ideas on this method, we describe several extensions: quantitative trait studies; and case-control studies with haplotypes and multiallelic markers. For each study design our goal is to achieve control similar to that obtained for a family-based study, but with the convenience found in a population-based design.

1 Introduction

Five decades ago Sewell Wright (1951) introduced the formal concepts underlying what geneticists term population structure. Wright argued that non-random and particularly local mating, together with genetic drift, tended to form genetic subpopulations within populations. Wright's substructure was a statistical concept, as it would be impossible to delineate subpopulations except in terms of the greater probability of relatedness of individuals drawn from within as opposed to between subpopulations. The significance of substructure, then as now, was its impact on evolutionary processes, although the significance of substructure for evolutionary processes was not and is not universally recognized (Fisher and Ford 1950; Fisher 1953; Nei 1987).

During 1972, Ci Ci Li underscored for geneticists the importance of population substructure for an entirely different subject, the discovery of gene-disease associations through the analysis of population samples. Li developed the concept for a particular case, analagous to inbreeding, but his results extend easily to the more-widely known case. Specifically, for two subpopulations with allele frequency p_1 and p_2 at one biallelic locus and q_1 and q_2 at a second locus, there will be a statistical association among genotypes in a sample from the whole population, whether the two loci were linked or not, as long as $p_1 \neq p_2$ and $q_1 \neq q_2$. The degree of association is a function of the variances and covariances of alleles in the subpopulations, as well as the fraction of the sample drawn from each subpopulation. While these observations had population genetic significance, Li highlighted its importance for genetic association studies when disease and marker allele frequencies varied among populations. Li's observations were important because it was popular to relate the plethora of human diseases to the handful of genetic markers that were available at the time. Most of the studies used a case-control design, a special form of population sample in which 'cases' or diseased individuals were oversampled relative to their frequency in the population. While many studies yielded negative results, a surprising number of genetic diseases were found to be partly attributable to these handful of markers, especially those associated with the HLA region on chromosome 6p (Thomson 1988; Risch 2000).

The explosion in the development of genetic markers since the mid 1980's tended to change genetic epidemiolgical designs. Linkage analyses became prominent, particularly

whole genome analyses (Ott 1991), and continue to this day. The success enjoyed by linkage methods for Mendelian diseases – those for which the inheritance patterns within families make a genetic model relatively obvious – has been nothing less than phenomenal. Gene mutations underlying hundreds of simple genetic diseases have been uncovered, and it is safe to say few such mutations will remain unknown a few years hence. For complex diseases – those for which there is clearly a genetic basis, but for which the inheritance pattern is fuzzy and the genetic model obscure – the picture is not so rosy. Despite substantial effort during the past decade on complex diseases, such as schizophrenia, the yield of disease-causing mutations has been small: none, in fact, for some common mental disorders.

It now appears that the genetic variants relevant for most complex diseases have subtle effect on liability, and hence most genetic linkage studies to date were underpowered. A possible solution is to collect larger samples and implement more thoughtful linkage analyses (Blangero *et al.,* 2001); another solution is a back-to-the-future approach, a return to association studies (Risch and Merikangas 1996). Evolutionary biologists will recognize immediately that genetic association designs are only useful when the evolutionary processes underlying the disease cause diseased individual to be more closely related than individuals chosen at random from the population would be. This differential in degree-of-relatedness, along with other design features, determines the power of the association analysis. Pure linkage analysis, on the other hand, ignores 'evolutionary' relatedness, using only the distribution of phenotypes and genotypes within narrowly-delimited families to determine the location of disease genes (Ott 1991). Thus, under certain conditions, which are more easily defined in theory than determined in practice, association analyses can have far greater power than linkage analysis to determine the genetic underpinnings of complex disease (Risch and Merikangas 1996; Knapp 1999a; Abel and Muller-Myhsok 1998).

Genetic association analysis also became more sophisticated in recent times. Bootstrapping off the widespread collection of families for whole genome linkage analysis, especially the recruitment of affected sib-pairs and their parents, Spielman *et al.,* (1993) introduced the TDT test, a simultaneous test of linkage and association. By testing both linkage and association using the transmission of alleles from parents to offspring, this 'family-based' approach completely obviates concerns about population substructure.

Since Spielman et al.'s 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 (e.g., Allison 1997; Spielman and Ewens 1998; Boehnke and Langefeld 1998; Knapp 1999b; Rabinowitz and Laird 2000; Horvath *et al.,* 2000; Sinsheimer *et al.,* 2000; Zhu and Elston 2001; Seltman *et al.,* 2001). In fact, the TDT and allied tests, based on the recruitment of partial or entire nuclear families, has essentially supplanted case-control as the method of choice for genetic epidemiological studies.

The degree of success of family-based studies could not have been anticipated by its original proponents, and cogent arguments can be made that its success has been unfortunate (Risch and Teng 1998; Morton and Collins 1998). As Risch and Teng (1998) note, if individual genes have only a subtle impact on liability for complex disease, then large samples will be required to identify even some of the variants affecting liability. Family-based samples, by design, mitigate against large samples because of the difficulty and expense of collecting families. And, even for the same sample sizes, TDT is less powerful than case-control 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 facile and powerful case-control designs as an initial screen for liability genes, with more controlled family-based follow-up studies. Indeed Spielman *et al.,* (1993), the orginal proponents of the TDT, suggested just such an approach. Still, the reality of population stratification and the false positives it generates weighs against case-control designs.

Mulling over this rather interesting problem from our perspective, namely evolutionary genetics, genetic epidemiology and statistics, we thougt there might be an alternative path. Our thinking was aided by the revolution in molecular tools, especially the promise of massive, inexpensive genotying. We reasoned it would be possible to account for the impact of substructure by using the distribution of markers in the sampled genomes – what we called Genomic Control or GC (Devlin and Roeder 1999). While it comes at the obvious price of additional genotyping, specifically genotyping multiple loci unlikely to affect liability, GC opens up the possibility of using population-based samples and controlling the false positive rate.

For example, for a case-control analysis of candidate genes, one GC approach computes chi-square 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 by the impact of population stratification and cryptic relatedness, 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 are not detected from null loci, then GC is identical to a standard test of independence for a case-control design.

Since we introduced the original concept of GC, we and others have developed the approach in various ways. In this review, we outline the basic theory behind GC, describe the various implementations to date, and then discuss some open questions regarding this methodology.

2 Confounding

The case-control design oversamples 'affected' individuals from the population and contrasts this sample with an undersampled set of controls, with the goal to determine if a particular variable, such as counts of alleles at a locus, differ between the two samples. Unobserved variables, such as membership in subpopulations, can create spurious correlations between variables or confounding. Confounding can have two effects on test statistics; they can be biased and/or overdispersed. While bias can be a critical factor for traditional epidemiological studies, we argue that overdispersion is the dominant consequence of confounding in genetic studies.

In this section, we explore the impact of confounding due to population substructure via a case-control study to assess association between alleles at a biallelic disease and a linked or unlinked locus. To do so we define several random variables that reflect the basic elements of a genetic study, albeit at a simplified level to facilitate exposition. Let C be an unobservable indicator of subpopulation membership, $c = 1, \ldots, m$; Y be a binary indicator of disease status; X be a disease susceptability gene; and G be genotype, which is reduced to two levels (A, a). This genotype notation is natural for recessive and dominant models for effects of alleles on disease liability, but also applies to additive models if alleles are treated as observations.

2.1 Bias in Case-Control Studies

In a case-control setting the response is genotype and the covariate is disease status, Y. We define δ to be the case-control effect

$$\delta = P(G = A|Y = 1) - P(G = A|Y = 0). \tag{1}$$

Under the null hypothesis δ is non-zero solely due to bias.

Figure 1 illustrates the situation of no confounding. The arrow from X to Y denotes a causal relationship, and the circle around X indicates X is not directly observable. The first graph in Figure 1 shows the null hypothesis that G and X are unlinked. The second graph has a double-headed arrow between X and G to indicate that they are linked and associated. (Technically, this is called a mixed ancestral graph (Richardson and Spirtes 2001) and the double-headed arrow actually represents the unobservable history that created the association between X and G, but those details are unnecessary here.) In this case, Y and G are associated if, and only if, X and G are associated. Since Y and G are observable, this provides a way to check for linkage and association between G and the unobservable X(ignoring the possibility of strong selection).

The situation is more complicated in the presence of substructure. Figure 2 shows the graphs for this case with C representing subgroups. We initially assume that the environmental component of the disease is small relative to the genetic component; hence there is no arrow from C to Y. In this case, under the null hypothesis of no association between X and G, it can be shown that G and Y are uncorrelated, given C, but that they are correlated marginally. In other words, $\delta \neq 0$ even under the null hypothesis.

To model the effect of subpopulation stratification, let $p_c = P(G = A|C = c)$, $r_c = P(Y = 1|C = c)$, and $w_c = P(C = c)$. It follows that

$$P(C = c|Y = 1) = \frac{r_c w_c}{\sum_l r_l w_l},$$

and

$$P(C = c|Y = 0) = \frac{(1 - r_c)w_c}{\sum_l (1 - r_l)w_l}. \tag{2}$$

Assume that G and Y are independent, conditional on C, under the null hypothesis. Let $d_c = P(C = c|Y = 1) - P(C = c|Y = 0)$. It follows from equation (1) that,

$$\delta = \sum_c p_c \left[P(C = c|Y = 1) - P(C = c|Y = 0)\right] = \sum_c p_c d_c. \tag{3}$$

In general δ may be negative or positive for any locus under study, even under the null hypothesis. Moreover, δ does not decrease as the sample size increases. In the simplest case, if there were only two subpopulations and both genotype A and disease ($Y = 1$) were more prevalent in one subpopulation than the other, then δ would be positive.

Next we seek to understand the behavior of δ. We focus on a test statistic based on an additive genetic model (i.e., G represents a single allele). We assume a substructured population with allelic correlation defined by F_{st} (Wright 1969). Assume $p_c = P(A|C = c)$ is an i.i.d. random variable, $c = 1, \ldots, m$, with mean p and variance $F_{st}p(1-p)$. It follows that $E[\delta] = p\sum_c d_c = 0$; i.e., in expectation (averaged over randomly selected subpopulations) the case-control bias has mean zero. The variance of the bias is

$$E[\delta^2] = \mathrm{Var}\left[\sum_c d_c p_c\right] = F_{st}p(1-p)\sum_c d_c^2. \tag{4}$$

This quantity is zero if there is no variability in allele frequencies due to population substructure.

In the remainder of this subsection we investigate $E[\delta^2]$ under a broad range of conditions and find that, under certain conditions, it is of modest size even in the presence of extreme population substructure. A key fact to note is that, regardless of the magnitude, the variance of the bias is always proportional to $p(1-p)$, where p is the allele frequency under investigation. This fact will be important for the GC method described shortly because bias due to genetic or environmental heterogeneity across subpopulations can be estimated.

Consider a rare, dominant Mendelian disorder for which any individual with at least one copy of the deleterious allele has the disease with probability one (see Fig. 3). In this simple setting, r_c is the probability an individual has at least one copy of the deleterious allele, which is approximately equal to twice the frequency of the allele in the c'th subpopulation, $2q_c$. Assume that q_c varies across subpopulations with mean $q = 0.0005$ and variance $F_{st}q(1-q)$. The distribution of r_c across subpopulations follows directly. Allele frequencies for A also vary across subpopulations and this determines the distribution of p_c for the locus under investigation. As with the disease gene, we consider a model of variability across subpopulations determined by F_{st}. Let $p = P(A)$ be the marginal probability of A, with the variance across subpopulations equal to $F_{st}p(1-p)$. Setting $F_{st} = 0.01$, we can investigate how large the bias might be and how it varies as a function

of the allele frequencies of the genotype under study. Consider a sample of markers with allele frequencies uniformly distributed in the interval $[0.05, 0.50]$ and sample from *two subpopulations of equal size*. For each of 1000 samples, we compute the bias of the test statistic. From this experiment (Fig. 4), two points are evident: (i) the absolute bias can be sizable; and, (ii), the variance of the bias is proportional to $p(1-p)$. This experiment provides a partial explanation for why association tests have fallen into disfavor.

By contrast, consider a complex disorder with an additive model for disease phenotypes, and $X_{i1}, \ldots, X_{iL}$ denoting the number of liability alleles possessed by the i'th individual at locus $l, l = 1, \ldots, L$. Let $q_{cl} = \frac{1}{2}E(X_{i\ell}|C = c)$ be the frequency of liability alleles at locus l. Assume for a moment that the combined effect of environmental contributions to disease susceptibility do not differ strongly across subpopulations (Fig. 5).

Regarding the r_c's as random and from the definition of d_c, the following is true: $\mathrm{Var}[d_c] \to 0$ as $\mathrm{Var}[r_c] \to 0$. Furthermore, δ converges in probability to 0 as $\mathrm{Var}[d_c] \to 0$. It follows that δ will be near zero if r_c does not vary appreciably across subpopulations. We look to understand the conditions that determine the size of δ. Consider a logistic model for disease susceptability

$$P(Y = 1|X_1, \ldots, X_L) = \frac{e^{\beta_0 + \sum_{\ell=1}^{L} \beta_\ell X_{i\ell}}}{1 + e^{\beta_0 + \sum_{\ell=1}^{L} \beta_\ell X_{i\ell}}}.$$

Note that the distribution of $Y|X_1, \ldots, X_L$ equals the distribution of $Y|X_1, \ldots, X_L, C$. Now,

$$\begin{aligned} r_c &= E[P(Y = 1|X_1, \ldots, X_L, C = c] \\ &= E\left(\frac{e^{\beta_0 + \sum_{\ell=1}^{L} \beta_\ell X_{i\ell}}}{1 + e^{\beta_0 + \sum_{\ell=1}^{L} \beta_\ell X_{i\ell}}}\right) \\ &\approx \frac{e^{\beta_0 + 2\sum_{\ell=1}^{L} \beta_\ell E(X_{i\ell}|C=c)}}{1 + e^{\beta_0 + 2\sum_{\ell=1}^{L} \beta_\ell E(X_{i\ell}|C=c)}} \\ &= \frac{e^{\beta_0 + 2\sum_{\ell=1}^{L} \beta_\ell q_{c\ell}}}{1 + e^{\beta_0 + 2\sum_{\ell=1}^{L} \beta_\ell q_{c\ell}}}. \end{aligned}$$

For each locus l, assume that q_{cl} is a random variable with mean q_l and variance $F_{st} q_l(1 - q_l)$. Now consider the asymptotics in which, as L grows, the total effects of the disease genes remains bounded and no one gene dominates. This implies that $||\beta|| = O(1)$

and $\beta_\ell = O(1/L)$ where $||\beta|| = \sqrt{\sum_\ell \beta_\ell^2}$. Taking the following expectation over this distribution, it follows from the delta method that, for large L,

$$\begin{aligned} E(r_c) &= \frac{e^{\beta_0+2\sum_{\ell=1}^{L}\beta_\ell E(q_{c\ell})}}{1+e^{\beta_0+\sum_{\ell=1}^{L}\beta_\ell E(q_{c\ell})}} + O\left(L^{-1/2}\right) \\ &= \frac{e^{\beta_0+2\sum_{\ell=1}^{L}\beta_\ell q_\ell}}{1+e^{\beta_0+\sum_{\ell=1}^{L}\beta_\ell q_\ell}} + O\left(L^{-1/2}\right) = O(1). \end{aligned}$$

Because $\beta_\ell^2 = O(L^{-2})$, $\sum_{\ell=1}^{L} q_\ell(1-q_\ell)\beta_\ell^2 = O\left(L^{-1}\right)$,

$$\begin{aligned} \mathrm{Var}[r_c] &= \frac{1}{(1+e^{\beta_0+2\sum_{\ell=1}^{L}\beta_\ell q_\ell})^2} Var(\beta_0 + 2\sum_{\ell=1}^{L}\beta_l q_{cl}) + O\left(L^{-1}\right) \\ &= \frac{1}{(1+e^{\beta_0+2\sum_{\ell=1}^{L}\beta_\ell q_\ell})^2} 4F_{st}\sum_{\ell=1}^{L} q_\ell(1-q_\ell)\beta_\ell^2 + O\left(\frac{1}{L}\right) \\ &= O\left(L^{-1}\right) + O\left(L^{-1}\right) = O\left(L^{-1}\right). \end{aligned}$$

Consequently we can conclude that the bias is small for a complex disease when multiple liability loci contribute to the probability of disease, and the total effect of the environmental contributions to liability do not vary substantially across subpopulations.

While the conditions we have explored for obtaining a small bias are sufficient, they are by no means necessary. Other factors, such as many subpopulations each varying by a small amount, can also lead to small bias terms. For instance, Wacholder *et al.,* (2000) investigated the likely size of the bias for cancer studies, both empirically and theoretically. They found that the bias is likely to be small, and it decreases as a function of the number of subpopulations involved.

In the description of a complex disorder above, it is implicitly assumed that the environmental factors affecting susceptibility to disease are constant across subpopulations. This need not be true in reality. Key environmental covariates can vary by culture and hence by subpopulation. Such variability can create bias terms approaching the maximum possible for a given level of substructure. Moreover, these environmental effects also may interact with liability loci. However, these environmental effects do not affect the allele frequency distribution of "null" loci. The key feature required in the upcoming section is that $\sum_c d_c^2$ is constant across the genome for null loci and this is true regardless of whether the disease is primarily heritable or not.

In summary, we conclude that the bias term is likely to be small for complex diseases unless there are strong, subpopulation-specific environmental effects. Regardless of the

size of the bias, its variance $E[\delta^2]$ is still proportional to $p(1-p)$. In other words, the unknown contribution to bias from any environmental effects are constant across the null regions of the genome.

2.2 Variance in Case-Control Studies

In a substructured population the proportion of genotypes AA, aa and Aa in the population are described by $F_{st}p+(1-F_{st})p^2$, $F_{st}(1-p)+(1-F_{st})(1-p)^2$ and $2(1-F_{st})p(1-p)$, respectively. 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. More troublesome, however, is the fact that the allelic correlation extends across individuals from the same subpopulation. 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 among the observations from a common subpopulation, the usual statistical test for association can result in a rate of false positives exceeding the nominal level.

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. In addition, let $\zeta_c = P(C=c|Y=1)$ and $\omega_c = P(C=c|Y=1)$ denote the expected sample size of cases and controls from each of the $c = 1,\ldots,m$ subpopulations. The variance of $T = \sum_i X_i - \sum_j Y_j$ is highly dependent on the similarity between ζ_c and ω_c;

$$\begin{aligned}\mathrm{Var}(T) &= \sum_{i=1}^{N}\mathrm{Var}(X_i) + \sum_{j=1}^{N}\mathrm{Var}(Y_j) \\ &+ 2\sum_{i<l}\mathrm{Cov}(X_i,X_l) + 2\sum_{j<l}\mathrm{Cov}(Y_j,Y_l) \\ &- 2\sum_i\sum_j \mathrm{Cov}(X_i,Y_j).\end{aligned}$$

From above we have $\mathrm{Var}(X_i) = \mathrm{Var}(Y_j) = 2p(1-p)(1+F_{st})$. For any pair of genotypes from the same subpopulation

$$\mathrm{Cov}(X_i,X_l) = \mathrm{Cov}(Y_j,Y_l) = \mathrm{Cov}(X_i,Y_j) = 4F_{st}p(1-p), \quad i\neq l,\ j\neq l.$$

It follows that the variance of $\hat{\delta} = T/2N$ equals

$$4Np(1-p)\left[1+F_{st}+NF_{st}\sum_c\{\zeta_c(\zeta_c-1)+\omega_c(\omega_c-1)-2\zeta_c\omega_c\}\right], \tag{5}$$

in which $\sigma^2 = p(1-p)/N$.

The most extreme effect of substructure occurs if the cases and controls are drawn from two distinct subpopulations. In this instance even small values of F_{st} can have a large impact on the distribution of T. Alternatively, the variance is minimized when disease status is independent of subpopulation membership. In this scenario population admixture has essentially no impact on the distribution of the test statistic. Arguably the situation most frequently encountered in practice is one for which the probability of disease varies somewhat by subpopulation.

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. We generalize (5) to incorporate cryptic relatedness. F_{st} is the probability that uniting gametes are identical-by-descent or *ibd* when they are drawn from the same subpopulation. The kinship coefficient, f_{ij}, gives a related quantity: for relatives i and j, it is the probability that an allele selected randomly from i and a allele selected randomly from the same autosomal gene of j are *ibd*. Both F_{st} and f_{ij} can be interpreted as the correlation between alleles. In fact, if i and j are related only because they are in the same subpopulation, then $f_{ij} = F_{st}$ and the following equation reduces to (5).

Define f_{ij}^X, f_{ij}^Y and f_{ij}^{XY} as the kinship coefficient between cases, controls, and cases and controls, respectively. Under the null hypothesis of no genetic association,

$$\begin{aligned} \mathrm{Var}[\hat{\delta}] &= \sigma^2 \times \left\{ 1 + F_{st} + \frac{2}{N}\sum_{i<j}^{N} f_{ij}^X + \frac{2}{N}\sum_{i<j}^{N} f_{ij}^Y - \frac{2}{N}\sum_{j=1}^{N}\sum_{i\neq j}^{N} f_{ij}^{XY} \right\} \\ &= \sigma^2 \times \tau^2, \end{aligned} \tag{6}$$

where the first term is $\mathrm{Var}[\hat{\delta}]$ assuming independent samples and τ^2 is the inflation of variance due to correlated alleles. Because τ^2 increases as a function of N it can be sizable even for small allelic correlations.

2.3 Distribution Theory

From the results in the previous subsections it follows that, given δ, $\hat{\delta}/\sigma \approx N(\delta, \tau^2)$. Hence, given δ, $(\hat{\delta}/\sigma)^2 \sim \tau^2\chi_1^2(\delta^2)$. But now consider multiple markers and let $\delta_k = P(G_k = A_k|Y=1) - P(G_k = A_k|Y=0)$ for marker k. Define $\hat{\delta}_k$ and σ_k^2 analogously. Hence, given δ_k, $\hat{\delta}_k \sim N(\delta_k, \tau^2)$. Now we investigate the marginal distribution of the

$(\hat{\delta}_k/\sigma_k)^2$ over the markers. Recall that $\delta_k = \sum_c p_{ck} d_c$. If the number of subpopulations is not too small, the central limit theorem implies that $\delta_k \approx N(0, V_k \sum_c d_c^2)$ where $V_k = Var(p_{ck}) = F_{st} p_k (1 - p_k)$. Note that $\eta^2 = V_k \sum_c d_c^2 / \sigma_k^2 = \frac{N F_{st}}{4} \sum_c d_c^2$ Hence, we can write $\delta_k \stackrel{d}{=} (V_k \sum_c d_c^2)^{1/2} W$ where $W \sim N(0,1)$. Also, we can write $\hat{\delta}_k \stackrel{d}{=} \delta_k + \sigma_k Z$ where $Z \sim N(0, \tau^2)$. Thus we have

$$\begin{aligned} \left(\frac{\hat{\delta}_k}{\sigma_k}\right)^2 &\stackrel{d}{=} \left(\frac{\delta_k + \sigma_k Z}{\sigma_k}\right)^2 \\ &= \left[N\left(0, \eta^2 + \tau^2\right)\right]^2 \\ &\stackrel{d}{=} (\eta^2 + \tau^2)\chi_1^2. \end{aligned}$$

From this discussion we see that confounding can lead to bias and overdispersion, both of which can produce excess false positives when testing for association, especially when N is large.

3 Genomic Control

During the past few sections, we developed an intriguing connundrum. The impact of most genes on complex disorders appears to be subtle, making the collection of large samples highly desirable. Large, family-based samples, which are immune to the impact of substructure, can be prohibitive in terms of time, money, and number of willing participants and this fact motivates recruitment of population-based samples, such as case-control data. Yet the impact of substructure and more direct relatedness on population-based studies is to raise the false-positive rate, making it more difficult to separate the wheat from the chaff in genetic studies.

An early approach to evaluating the results of population-based studies was to test the samples for violations of Hardy-Weinberg (HW) 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? The weaknesses of these approaches, however, can be ameliorated

by using methods such as those developed by Devlin and Roeder (1999) and Pritchard and colleagues (2000).

In 1999, we proposed an alternative approach – Genomic control or GC. Building on standard results from evolutionary theory [e.g. Wright (1969), Lewontin and Krakaurer (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.

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. Since the GC concept was introduced, the Structured Association or SA method has been developed. Pritchard *et al.,* (2000a) proposed using marker 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 (see also Pritchard *et al.,* 2000b); in the next stage, a test of association is conducted within subpopulations. Pritchard *et al.'s* (2000a) work, which falls under the rubric of latent class models (the subpopulations are the unknown, latent classes), was taken further in recent work by Satten *et al.,* (2001). For related work, see Schork et al. (2001), who applied this general idea to 44 microsatellite loci used in a renal failure study.

In what follows, we focus on the GC approach. For detailed development of the SA models, see Pritchard and Donnelly (2001) in this volume.

We initially illustrate the GC approach for a biallelic marker and a case-control sample. We then follow with an outline of how these results extend to some other experimental designs. In each scenario population substructure introduces a bias and a variance inflation factor, neither of which can be directly estimated for any particular locus. However, because these nuisance parameters are approximately constant across the genome, they can be estimated provided numerous 'null' loci are sampled.

Before we delve too deeply into the GC approach, we reiterate that this approach will only work when unknown constants, such as τ^2 and η^2, are approximately constant for

all null markers. Importantly, for the model of cryptic relatedness, the variance inflation is due to correlations or kinship coefficients unrelated to properties of individual loci. Therefore, under this model, the variance inflation is the same for all markers throughout the genome. For neutral alleles and equal mutation rates across loci, theory suggests F_{st} is constant for alleles both within and between loci regardless of their frequencies (e.g., Wright 1969). Moreover, the variance in F_{st} is minimized if the subjects are drawn from the same ethnic group [cf. Lewontin and Krakauer (1973) with Robertson (1975).] In real populations, F_{st} does vary. Furthermore, it can vary as a function of allele frequencies, depending on the populations examined. For example, when we fit a line to the data on F_{st} and allele frequencies for worldwide populations reported in Cavalli-Sforza *et al.,* (1994), we found a significant relationship (Bacanu *et al.,* 2000, Fig. 5); however, the greatest change occurred for $p < 0.1$, and F_{st} did not vary with p thereafter. For a set of European populations, there was no relationship between F_{st} and p (Bacanu *et al.,* 2000, Fig. 5), and the variability of F_{st} was small.

3.1 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 χ^2 test based upon the 2×2 allelic table. This statistic, S_k, is directly related to the statistic discussed previously: $S_k = \left(\hat{\delta_k}/\hat{\sigma_k}\right)^2$. Under the null hypothesis, for large N, S_k is approximately distributed as a scaled, χ^2_1 random variable with scaling parameter $\lambda = \eta^2 + \tau^2$. Provided τ^2 and η^2 are constant across the genome, λ can be estimated. Thus tests for association can be adjusted by dividing S_k by $\hat{\lambda}$. (see Devlin and Roeder 1999; Bacanu *et al.* 2000).

To estimate λ two choices are natural: a robust estimator such as the median of the χ^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 Goldstien (2001) recommend a more conservative correction to account for the sampling variability in $\hat{\lambda}$; in practice this leads to estimates for λ 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.

3.2 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. To test if a single locus is associated with the phenotype, we work with the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i, \tag{7}$$

where X_{i1} is the number of A alleles minus the expected number in the i'th individual. We test whether the slope is different from zero. Define $\sigma^2 = \mathrm{Var}[Y_i]$ and let $\rho = \mathrm{Cov}[Y_i, Y_j]$ denote the covariance of phenotypes of individuals in the same subpopulation. Of the $N(N-1)/2$ pairs of individuals in the study, let R denote the number of pairs with positive covariance.

The usual estimator of the parameter of interest is $\hat{\beta}_1$. Two factors perturb the distribution of $\hat{\beta}_1$ from that expected in the typical regression setting. (i) Due to positive correlation among subjects within a subpopulation, the variance is increased over that expected under the independence model. And, (ii), due to population substructure $E[\hat{\beta}_1]$ is not equal to zero under the null hypothesis.

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. It can be shown that the actual variance of $\hat{\beta}_1$ is

$$\begin{aligned} \mathrm{Var}[\hat{\beta}_1] &\approx \frac{\sigma^2}{2N(1+F_{st})p_1(1-p_1)}\left[1 + \frac{4RF_{st}\rho}{N(1+F_{st})\sigma^2}\right] \\ &= SE^2_{ind}[\hat{\beta}_1] \times \tau^2; \end{aligned}$$

see Devlin *et al.,* (2001) for details. The adjustment to the usual variance term, τ^2, is the inflation factor due to correlation among the subjects in the study. If $F_{st} = 0$, then $\tau^2 = 1$ and the variance reduces to $\sigma^2/\{2Np_1(1-p_1)\}$, which is estimated by $SE^2_{ind}[\hat{\beta}_1]$. If R and F are large, then τ^2 can be substantial. Alternatively, if there are many small subpopulations, then R will be small and the impact of population substructure on the variance will be small.

As discussed in the section on confounding, we also anticipate some bias in this test so that $E[\hat{\beta}_1] = \delta \neq 0$, even under the null hypothesis. Using arguments analogous to

those used to obtain $E[\delta^2]$ for case-control samples, we can show that: (i) $E[\delta] = 0$; (ii)Var$[\delta] = \eta^2 SE^2_{ind}[\hat{\beta}_1]$, where η^2 is a constant function across the genome; and (iii) the bias is typically small unless there is a strong subpopulation-specific environmental effect. Consequently, as in the case-control setting, the test statistic $QT = (\hat{\beta}_1/SE_{ind}(\hat{\beta}))^2$ is approximately distributed $(\eta^2 + \tau^2)\chi^2_1$.

For any single locus, $\lambda = (\eta^2 + \tau^2)$ cannot be directly estimated. It depends upon unknown allele frequencies and allelic effects at unspecified loci that drive the heritability of the trait. However, both constants are approximately constant regardless of which loci are under study. 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.

3.3 GC for Case-control Studies with Multiallelic Markers and Haplotypes.

To have good power for modest sample sizes, the causal mutation must appear in combination with a relatively uncommon marker allele. Otherwise the increased risk associated with the marker will be difficult to detect. This suggests that association studies based upon multiallelic markers or haplotypes, which are treated as such, can be considerably more informative than biallelic markers. Ideally a single polymorphism is suspected of enhancing the risk of disease. However, when such a hypothesis is not available, the omnibus χ^2 test for association is the usual choice. This statistical test has many degrees of freedom and hence can have small power even for substantial associations. Furthermore, if susceptability alleles are not rare, there is no reason to expect strong associations with any single marker allele or haplotype. To combat this problem several authors have suggested measures of association that use a single degree of freedom (e.g., van der Meulen *et al.,* 1997, Bourgain *et al.,* 2001). Here we consider one candidate 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. Next assume there are L_k allele types for haplotype segment k, with corresponding allele relative frequency as $(p_{Y1(k)}, p_{Y2(k)}, ...p_{YL_k(k)})$ in cases and $(p_{X1(k)}, p_{X2(k)}, ...p_{XL_k(k)})$ in controls. The probability

of two case haplotype segments matching is then

$$\Pr(\text{Hap}_i \text{ and } \text{Hap}_j \text{ match}) = \sum_{l=1}^{L_K} p^2_{Yl(k)}.$$

A suitable test statistic could be based upon the difference in the matching probability:

$$T_k = \sum_{l=1}^{L_k} p^2_{Yl(k)} - \sum_{l=1}^{L_k} p^2_{Xl(k)}$$

This test is sensitive to detecting excess matching in case subjects versus control subjects and the sensitivity holds whether their are numerous small clusters or a few larger ones. For this reason we expect this statistic to perform well even when susceptability alleles are common.

For sufficiently large samples, T_k follows a normal distribution $N(\mu_k, \text{Var}(T_k))$. The null hypothesis of this test is the haplotype segment k is not proximate to the disease gene; the alternative hypothesis supposes the segment k is close to the disease gene, which produces extra "matchiness" in T_k. This statement can be summarized as $H_0 : \mu_k = \mu_0$ vs. $H_a : \mu_k > \mu_0$. Any T_k's substantially different from the alternative hypothesis are defined as "outliers". The first step is to standardize T_k under the null hypothesis by calculating the mean and variance of T_k to obtain $Z_k = (T_k - \mu_0)/\text{Var}(T_k)^{1/2}$.

Define $Y_{i(k)} = (Y^1_{i(k)}, Y^2_{i(k)}, ..., Y^{L_k}_{i(k)})$ as the multinomial coding of the ith case's allele type of haplotype segment k. For example, type 2 is coded as $(0, 1, 0, ..., 0)$, and type L_k is coded as $(0, 0, ..., 0, 1)$. Under the null hypothesis, $Y_{i(k)} = (Y^1_{i(k)}, Y^2_{i(k)}, ..., Y^{L_k}_{i(k)})$ and $X_{i(k)} = (X^1_{i(k)}, X^2_{i(k)}, ..., X^{L_k}_{i(k)})$ are identically, but not independently distributed as a Multinomial with sample size one and probability vector $(p_{1(k)}, ..., p_{L_k(k)})$.

The correlation between two individuals f_{ij} is equal to the probability of two genes are *ibd*; i.e., $\text{Corr}(Y^l_i, Y^l_j) = f_{ij}$. By the same reasoning, we obtain $\text{Cov}(Y^l_i, Y^h_j) = -p_l p_h f_{ij}$. Given these two correlations it is possible to show that $\text{Var}(T_k) = \tau^2 \sigma^2_k$, where τ^2 is of a form similar to that seen in equation (6) and σ^2_k is variance of T_k assuming all haplotypes are independent; details of these calculations are given in Tzeng *et al.,* (2001). Notice τ^2 is a constant over k, that is, no matter which haplotype segment we are considering, the correlation among individuals affects the variance of T_k multiplicatively in the same way. To implement the GC procedure, all we require are robust estimates of μ_0 and τ^2.

3.4 Detecting Outliers.

For all the methods described so far, we require an estimate of one or two parameters, such as λ or μ_0 and τ^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 loci tested will naturally be null, provided a robust estimator is chosen, the parameters can be estimated using all of the data. For more details, see Devlin and Roeder (1999) and Tzeng *et al.,* (2001).

For example, when the test statistics are distributed as $\lambda\chi^2_1(0)$, a robust estimator of λ is:

$$\hat{\lambda} = \{\text{median}(S_{c+1}, S_{c+2}, \ldots, S_n)/0.456\}.$$

More efficient estimators exist for λ, but the median provides a reliable estimate of the inflation factor even if a small fraction of the null loci actually affect liability to the disease or are linked to the gene under study. The effect of treating outliers as null loci in the estimation of $\hat{\lambda}$ is a slight positive bias in the estimator, which has the effect of decreasing both the power and the size of the test somewhat.

When examining g candidate genes, a Bonferroni correction provides the critical value for the multiple testing problem. The significant ones are the "outliers". Devlin and Roeder (1999) present a Bayesian procedure for performing GC for a genome scan which is more powerful, but also more complicated to implement. Tzeng *et al.,* (2001) presents a non-Bayesian method for outlier detection that is 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.

3.5 Methods for extremely dense markers.

Grant *et al.,* (1999) and Devlin *et al.,* (2000) present methods of genomic control for very fine scale data — either genome mismatch scanning data or genetic marker data providing such a dense grid of markers that identity-by-descent is essentially apparent. Both of these methods rely on comparing pairwise haplotypes of diseased individuals to

find regions with usually long segments shared among individuals who are not obviously related. The former determines significance based on a complex permutation test. The latter determines significance using a score test for U-statistics. Both methods are similar to the method of Tzeng *et al.,* (2001), described above, in that they do not restrict the association to be driven by a few ancestral haplotypes.

3.6 Concluding Remarks

Genomic Control (GC) is a new method 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 attempts to eliminate these spurious associations. In this way, we hope GC 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 (§3.1) and multiallelic (§3.3) loci , haplotypes (§3.3), and quantitative traits (§3.2). Future research, both from ours and other groups, will extend the potential applications of GC. For example, we have extended the methodology 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 (Roeder, Bacanu and Devlin, submitted).

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.

The relative performance of GC versus family-based association methods has been explored in Bacanu *et al.,* (2000). By comparing GC and TDT in the case-control setting, we conclude that GC is more powerful than TDT when population heterogeneity is like that of European populations. GC and TDT have about similar power under more extreme levels of substructure, such as a mixed sample of Caucasian and African Americans. Analyses by Ewens and Spielman (1995) show that TDT will outperform GC in analyses of admixed populations, such as Africans Americans. Analyses by Pritchard *et al.,* (2000) suggest that their SA models perform similarly to TDT in many settings. The relative performance of GC and SA approaches are compared in Pritchard and Donnelly (this issue).

Acknowledgements

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

References

Abel, L., and Muller-Myhsok, B., 1998. Maximum-likelihood expression of the transmission/disequilibrium test and power considerations. *Am J Hum Genet* **63**, 664-667.

Allison D.B., 1997. Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* **60**, 676-690.

Bacanu, S-A., Devlin, B., and Roeder, K., 2000. The power of genomic control. *Am J Hum Genet* **66**, 1933-1944.

Bacanu, S-A., Devlin, B., and Roeder, K., 2001. Association studies for quantitative traits in structured populations. *Genet Epidemiol.*

Benjamini, Y. and Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Statist Soc B* **57**, 289-300.

Blangero, J., Williams, J.T., and Almasy, L., 2001. Variance component methods for detecting domplex trait loci. *Adv Genet* **42**, 151-181.

Boehnke M., and Langefeld C.D., 1998. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am J Hum Genet* **62**, 950-961.

Bourgain, C., Genin, E., Holopainen, P., Mustalahti, K., Maki, M., Partanen, J., and Clerget-Darpoux, F., 2001. Use of closely related affected individuals for the genetic study of complex diseases in founder populations. *Am J Hum Genet* **68**, 154-159.

Cavalli-Sforza, L.L., Menozzi. P., and Piazza, A., 1994. "The History and Geography of Human Genes," Princeton University Press, Princeton, New Jersey.

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

Devlin, B., and Roeder, K., 1999. Genomic control for association studies. *Biometrics* **55**, 997-1004.

Devlin, B., Roeder, K., and Wasserman, L., 2000. Genomic control for association studies: A semiparametric test to detect excess haplotype-sharing. *Biostatistics* **1**, 369-387.

Devlin, B., Roeder, K., Otto, C., Tiobech, S., and Byerley W., 2001. Genome-wide distribution of linkage disequilibrium in the population of Palau and its implications

for gene flow in Remote Oceania, *Hum Genet* in press.

Ewens, W.J., and Spielman, R.S., 1995. The transmission/disequilibrium test: history, subdivision and admixture. *Am J Hum Genet* **57**, 455-464.

Fisher, R.A., 1953. Population Genetics. *Proc Roy Soc London*, Ser B **141**, 510-523.

Fisher, R.A., and Ford, E.B., 1950. The "Sewell Wright" effect. *Heredity* **4**, 117-119.

Grant, G.R., Manduchi, E. Cheung, V.G. and Ewens, W.J., 1999. Significance testing for direct identity-by-descent mapping. *Ann Hum Genet* **63**, 441-454.

Horvath, S., Laird, N.M., and Knapp, M., 2000. The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers. *Am J Hum Genet* **66**, 1161-1167.

Knapp, M., 1999a. A note on power approximations for the transmission/disequilibrium test. *Am J Hum Genet* **64**, 1177-1185.

Knapp, M., 1999b. The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/ disequilibrium test. *Am J Hum Genet* **64**, 861-870.

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. *Ann Hum Genet* **33**, 23-29.

Morton, N.E., and Collins, A., 1998. Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci USA* **95**, 11389-11393.

Nei M (1987) "Molecular Evolutionary Genetics," Columbia University Press, New York.

Ott, J., 1991. "Analysis of Human Genetic Linkage," The Johns Hopkins University Press, Baltimore.

Pritchard, J.K., and Rosenberg, N.A., 1999. Use of unlinked genetic markers to detect population stratification in association studies *Am J Hum Genet* **65** 220-228.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., and Donnelly, P., 2000a. Association mapping in structured populations. *Am J Hum Genet* **67**170-181.

Pritchard, J.K., Stephens, M., and Donnelly, P., 2000b. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.

Rabinowitz, D., and Laird, N., 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered* **50**, 211-223

Reich D.E., and Goldstein, D.B., 2001. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* **20**, 4-16.

Richardson, T. and Spirtes, P., 2001. "Ancestral Graph Markov Models," Technical Report no. 375. University of Washington Department of Statistics.

Risch, N.J., 2000. Searching for genetic determinants in the new millennium. *Nature* **405**, 847-856.

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 disease. I. DNA pooling. *Genome Res* **8**,1273-1288.

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

Satten, G.A., Flanders, W.D., and Yang Q., 2001. Accounting for Unmeasured Population Substructure in Case-Control Studies of Genetic Association Using a Novel Latent-Class Model. *Am J Hum Genet* **68**, 466-477.

Schork, N.J., Fallin, D., Thiel, B., Xu, X., Broeckel, U., Jacob, H.J., and Cohen, D., 2001. The future of genetic case-control studies. *Adv Genet* **42**, 191-212.

Seltman, H., Roeder, K., and Devlin, B., 2001. TDT meets MHA: family-based association analysis guided by the evolution of haplotypes. *Am J Hum Genet*, in press.

Sinsheimer, J.S., Blangero, J., and Lange, K., 2000. Gamete-competition models. *Am J Hum Genet* **66**, 1168-1172.

Spielman, R.S., McGinnis, R.E., and Ewens, W.J., 1993. Transmission test for linkage disequilibrium: the insulin gene region andinsulin-dependent Diabetes Mellitus (IDDM).

Am J Hum Genet **52**, 506-516.

Thomson, G., 1988. HLA disease association: Models for insulin-dependent diabetes Mellitus and the study of complex human genetic disorders. *Annu. Rev. Genet.* **22**, 31-50.

Tzeng, J-Y., Wasserman, L., Byerley, W., Devlin, B., and Roeder, K., 2001. Outlier detection and false discovery rates for whole-genome DNA matching. *J Amer Stat Assoc*, submitted.

van der Meulen, M.A., and Meermen, G.J., 1997. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Gen Epidemiol* **14**, 915-919.

Wacholder, S., Rothman, N., and Caporaso, N., 2000. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* **92**, 1151-1158.

Wright, S., 1969. "Evolution and the genetics of populations. Vol 2: The theory of gene frequencies," University of Chicago Press, Chicago

Wright, S., 1951. The genetical structure of populations. *Ann Eugen* **15**, 323-354.

Zhu, X. and Elston, R.C., 2001. Transmission/disequilibrium tests for quantitative traits. *Genet Epidemiol* **20**, 57-74.

Figure Legends

Figure 1. Relationship among variables Y, a binary indicator of disease status, X, a disease susceptability gene, and G, the genotype. Both figures represent the case of no confounding. The leading graphic displays the relationships under the null hypothesis (H_0), with the arrow from X to Y denoting a causal relationship. The trailing graphic displays the relationships under alternative hypotheses (H_1), with the double-headed arrow between X and G indicating that they are related.

Figure 2. Relationship among the variables depicted in Figure 1 in the presence of a confounding variable C. Confounding is due to subgroups. The circled variables are unobservable.

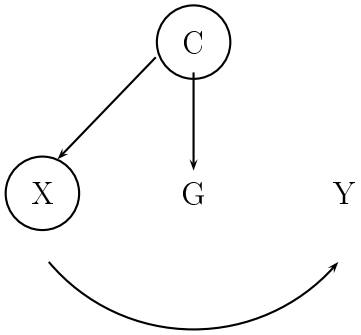
Figure 3. Relationship among the variables depicted in Figure 1 in the presence of a confounding variable C, with confounding due to subgroups, for a simple Mendelian disease. C is unobservable, p_c is the probability of drawing a genotype and r_c is the probability of being affected with the disease, both probabilities conditional on C. The leading graphic displays the relationships under the null hypothesis and the trailing graphic displays the relationships under the alternative hypothesis.

Figure 4. The Bias in a Case-Control study with a rare Mendelian Disease. We simulated an extreme case of population substructure: only two subpopulations and $F_{st} = 0.01$. The top panel shows the distribution of the absolute magnitude of the bias, $|\delta|$, for all values of p. The bottom panel shows $\sqrt{\delta^2/[p(1-p)]}$ as a function of p, the marker allele frequency.

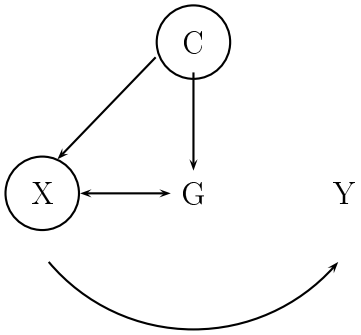
Figure 5. Relationship among variables Y, a binary indicator of disease status, disease susceptability genes $X_1, \ldots, X_L$ and genotypes $G_1, \ldots, G_k$ in the presence of a confounding variable C.

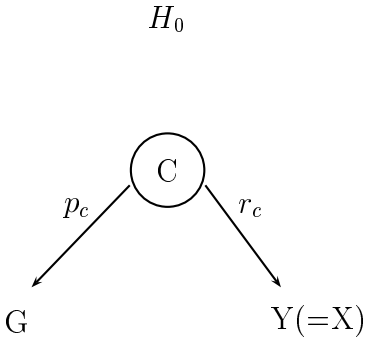

H_0
X
G
Y
H_1
X
G
Y

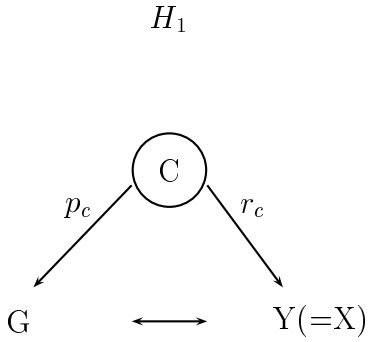
H_0

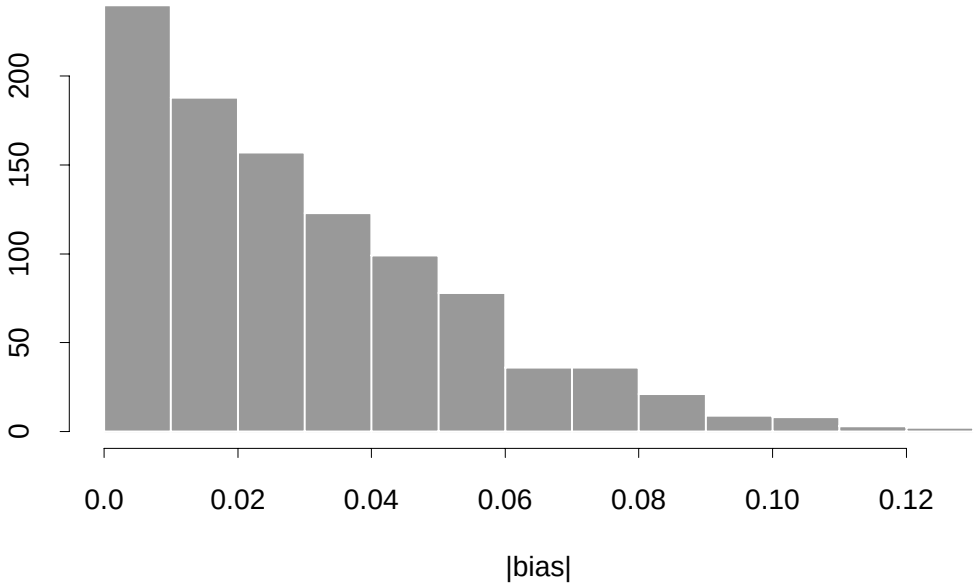


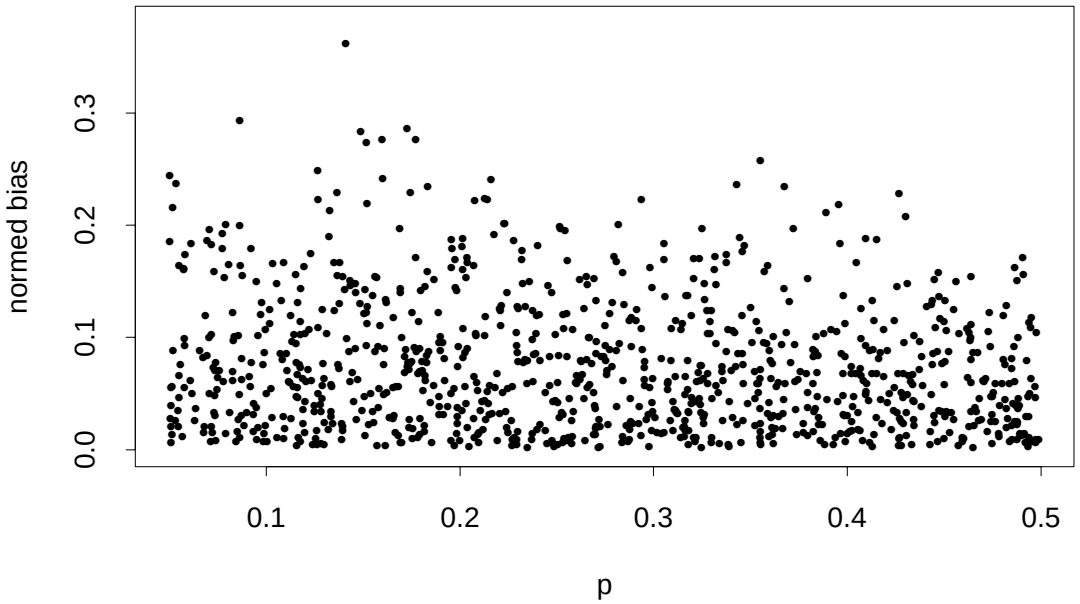
H_1

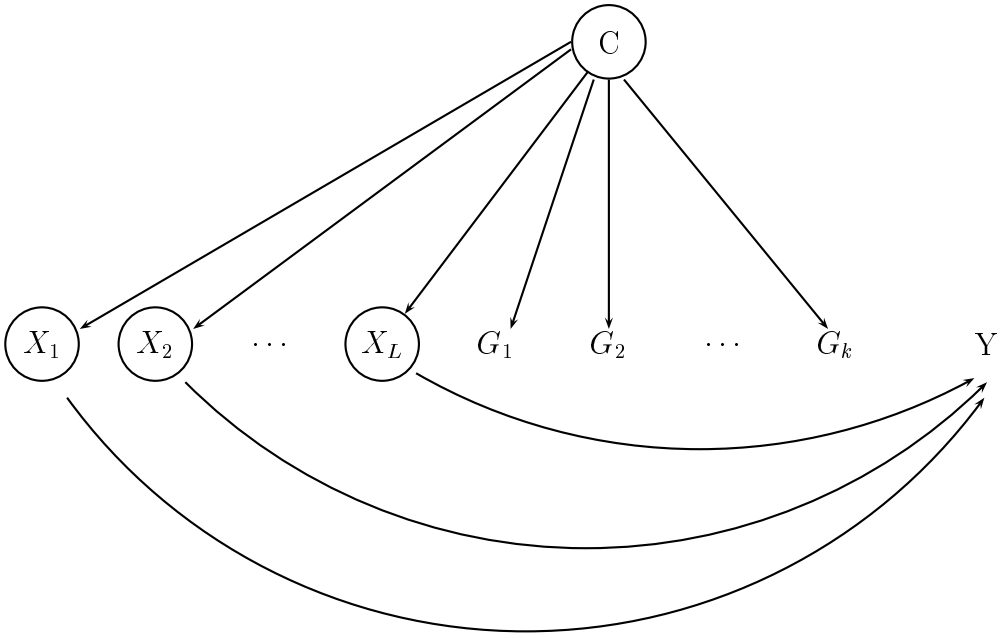



H_0
C
p_c
r_c
G
Y(=X)


H_1
C
p_c
r_c
G
Y(=X)


0
50
100
150
200
0.0
0.02
0.04
0.06
0.08
0.10
0.12
|bias|


normed bias
0.0
0.1
0.2
0.3
0.1
0.2
0.3
0.4
0.5
p


C
X_1
X_2
$\cdots$
X_L
G_1
G_2
$\cdots$
G_k
Y