Population Structure in Genetic Association Studies

Alice S. Whittemore
Stanford University School of Medicine
Department of Health Research and Policy
HRP Redwood Building, Room T204
Stanford, CA 94305-5405
E-mail: *alicesw@stanford.edu*

Abstract

Standard genetic association tests using case-control data are based on certain assumptions about the population from which study subjects were sampled. Two types of departure from these assumptions have been studied: population stratification and cryptic relatedness. Both types of departure have been called *population structure*. Each can lead to erroneous inferences due to differences between a test statistic's actual null distribution and the nominal one valid only for populations without structure. The differences can reflect either confounding bias or variance distortion. For each type of structure, adjusted test statistics have been proposed whose actual null distributions, in the presence of the structure, equal the nominal ones appropriate for unstructured populations. This paper reviews models for population stratificationand cryptic relatedness, and uses them to examine the effects of each on the Armitage trend test for case-control data. Specifically, population stratification can cause confounding bias but not variance distortion, while cryptic relatedness can cause variance distortion but not confounding bias. Consequently the adjusted statistics developed for population stratification (e.g. the latent variable methods of Pritchard et al. (1999, 2001); Satten et al. (2001); Schork et al. (2001); Wang et al. (2005)), address potential confounding bias but not variance distortion. Conversely, the adjusted statistics developed for cryptic relatedness (e.g. the Genomic Control (GC) methods of Devlin and Roeder (1999), Setakis et al. (2006) and Zheng et al. (2006)) address variance distortion but not confounding bias. These differences may explain the anomalous behavior of adjusted statistics when applied to populations with structure of a type that differs from the one for which the method was designed. They indicate that care is needed to specify the nature of the underlying structure anticipated for a given population, and to use appropriate methods to adjust for it.

Keywords: case-control studies, confounding bias, cryptic relatedness, genomic control, latent variables, population stratification, trend test, variance distortion

1 Introduction

Large-scale association studies offer substantial promise for unraveling the genetic basis of common human diseases. A problem with such studies is that the null distributions of standard test statistics may differ from their assumed nominal forms. This may occur when the population from which study subjects are sampled has some form of structure. Then incorrect inferences can occur due to discrepancies between the tests' actual and nominal type-1 error rates.

Two models for population structure have been proposed: a *population stratification* model and a *cryptic relatedness* model. Both assume that the population of interest can be decomposed into disjoint subpopulations whose memberships are unobserved. In the population stratification model, the unobserved subpopulations have different allele frequencies for a polymorphism of interest. If the subpopulations also differ in disease risk, inferences based on the full population are vulnerable to confounding bias. Several authors have proposed latent variable methods that produce adjusted test statistics having the nominal null distribution in the presence of confounding bias from population stratification (Pritchard et al., 1999, 2001; Satten et al., 2001; Schork et al., 2001; Wang et al., 2005). In the cryptic relatedness model, in contrast, the unobserved subpopulations have the same allele frequencies at the marker of interest; however their members share alleles IBD at any given marker. This IBD sharing induces correlation among the genotypes of different individuals, which in turn distorts the nominal null variance of test statistics. The Genomic Control (GC) methods provide adjusted test statistics having the nominal null distribution in the presence of such cryptic relatedness (Devlin and Roeder, 1999; Reich and Goldstein, 2001; Setakis et al., 2006; Zheng et al., 2006).

This paper reviews the population stratification and cryptic relatedness models. It shows that population stratification can cause confounding bias but not variance distortion, while cryptic relatedness can cause variance distortion but not confounding bias. Consequently the adjusted statistics using latent variable methods have the nominal null distribution in the presence of confounding bias due to population stratification, but not in the presence of variance distortion. Conversely, the adjusted statistics using GC methods have the nominal null distribution in the presence of variance distortion due to cryptic relatedness, but not in the presence of confounding bias. These differences may explain the erratic per-

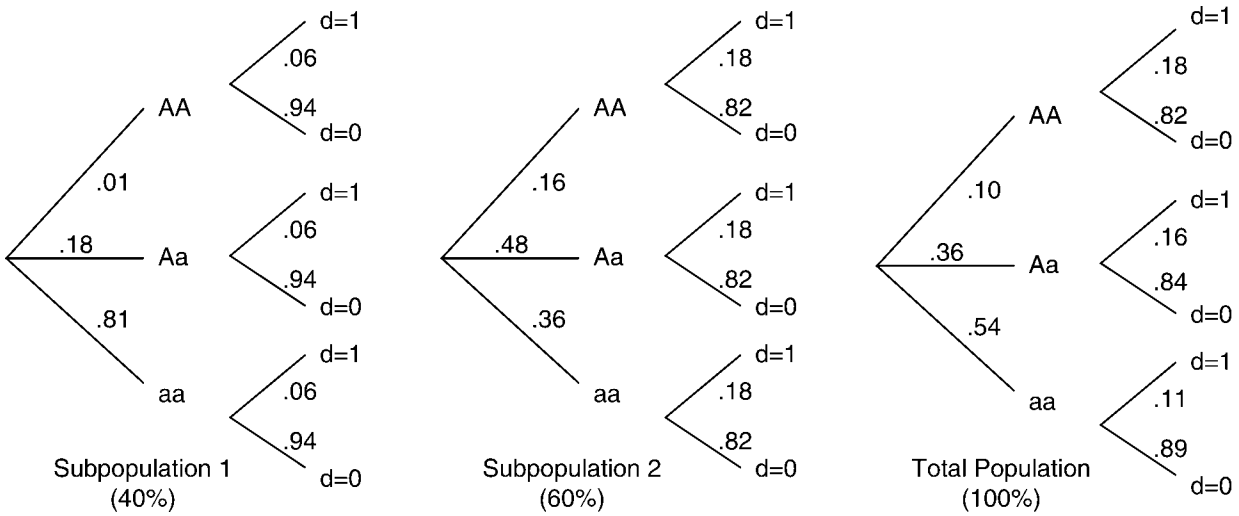

Figure 1: Distribution of genotypes and disease status among individuals in a hypothetical population containing 40% and 60% of its individuals in subpopulations 1 and 2, respectively.

formance of GC adjustment methods when applied to data simulated from stratified populations (Marchini et al., 2004; Campbell et al., 2005). The paper concludes with a discussion of diagnostics to assess the magnitude of both confounding bias and variance distortion.

2 The Population Stratification Model

A stratified population is defined to be one consisting of $J \geq 2$ subpopulations such that: 1) conditional on subpopulation membership, an individual's homologous marker alleles are independent of each other and of the alleles of other individuals; and 2) at least two subpopulations have different allele frequencies at one or more markers (Crow and Kimura, 1970, p 54; Elandt-Johnson, 1971, p 228).

To illustrate this type of structure, consider the subpopulations and genotypes shown in Figure 1. Here, the polymorphism of interest has two alleles, A and a, and the population consists of two subpopulations comprising 40% and 60% of the individuals. Let $y = 1$ if a chromosome bears allele A, with $y = 0$ otherwise, and let $g = 0, 1$ or 2 denote an individual's genotype (the sum of the y-values for his two homologous chromosomes). Given the subpopulations j and j' of two unrelated individuals, the indicators y for any pair of their chromosomes are assumed to be independent Bernoulli variables with means p_j and $p_{j'}$ and variances $p_j q_j$ and $p_{j'} q_{j'}$, respectively. Here $p_j = 1 - q_j$ is the frequency of allele A on chromosomes in subpopulation j, with $0 < p_j < 1$, $j = 1, \ldots, J$. (For the $J = 2$ subpopulations in Figure 1, $p_1 = 0.1$ and $p_2 = 0.4$.) The conditional distribution of an individual's genotype g, given his membership in subpopulation j, is binomial with parameters 2 and p_j. Its mean is $2p_j$ and its variance is $2p_j q_j$. Thus Hardy-Weinberg (HW) genotype frequencies hold within each subpopulation.

Although the conditional distribution of an individual's genotype, given his subpopulation, is binomial, the observed data allow inference only for its marginal distribution, averaged over the J subpopulations. This distribution is a mixture of J binomials. Its mean is

$$E[g] = 2\sum_{j=1}^{J} a_j p_j \equiv 2P, \tag{1}$$

where a_j is the proportion of the population in subpopulation j. Its variance is

$$\sigma^2 = \sum_{j=1}^{J} a_j (2p_j q_j) + 4\sum_{j=1}^{J} a_j p_j^2 - 4P^2 = 2PQ + 2s^2, \tag{2}$$

where

$$s^2 = \sum_{j=1}^{J} a_j p_j^2 - P^2. \tag{3}$$

In equation (3) s^2 represents both the inter-subpopulation variance of allele A, and the marginal covariance between alleles on the chromosomes of a single individual.

In summary, the population stratification model implies that the marginal distribution of one person's genotype is not binomial. (For the population in Figure 1 with $J = 2$, the genotype of an individual in the population has mean 0.56 and extra-binomial variance $\sigma^2 = 0.5032$.) However the genotypes of two unrelated individuals are marginally independent.

3 The Cryptic Relatedness Model

Cryptic relatedness occurs when a population consists of $J \geq 2$ subpopulations with the following properties: 1)

conditional on subpopulation membership, marker alleles of all individuals in the same subpopulation are correlated, with correlation coefficient $f > 0$; 2) marker alleles of individuals in different subpopulations are independent; and 3) all subpopulations have the same allele frequencies $p_j = p$ (Devlin and Roeder, 1999; Crow and Kimura, 1970, p 64; Elandt-Johnson 1971, p 213).

This model implies that within each subpopulation, an individual's genotype g is the sum of two correlated Bernoulli variables. Its conditional distribution, given the individual's membership in subpopulation j, is

$$\Pr(g|S=j) = \begin{cases} p^2 + fpq & g = 2 \\ 2pq - 2fpq & g = 1 \\ q^2 + fpq, & g = 0 \end{cases}, \quad j = 1, \ldots, J. \tag{4}$$

Since the distribution (4) is invariant across the subpopulations, it is also the marginal distribution of an individual's genotype. Its mean and variance are

$$E[g] = 2p, \tag{5}$$

and

$$\begin{aligned} \sigma^2 &= var(y_1 + y_2) = 2var(y_1) + 2cov(y_1, y_2) \\ &= 2pq + 2fpq = 2pq(1+f). \end{aligned} \tag{6}$$

The covariance of genotypes of two individuals, conditional on their membership in the same subpopulation, is $4fpq$. The marginal covariance of the genotypes g, g' of any two individuals is

$$cov(g, g') = \left(\sum_j a_j^2 \right) (4fpq). \tag{7}$$

To compare the two types of structure, we rewrite expression (2) as $\sigma^2 = 2PQ(1+F)$, where $F = s^2/PQ$. Then we compare the resulting expressions (1) and (2) and their parameters (P, Q, F) to expressions (5) and (6) and their parameters (p, q, f). We see that the marginal mean and variance of individual genotypes have the same forms under the two structure models, and that in both cases, the binomial variance of genotypes is inflated. However while genotypes of two individuals are uncorrelated under the population stratification model, they have positive covariance (7) under the cryptic relatedness model. Because of this difference, the two models have quite different implications for confounding bias and variance distortion.

4 Population Consequences

To investigate how the models for population stratification and cryptic relatedness affect the distribution of test statistics for a given population, we begin by reviewing the notions of confounding bias and variance distortion.

4.1 Confounding bias

An observational study of association between a disease and an exposure faces potential confounding by any attribute that is correlated with both disease and exposure (Kelsey et al., 1996, p 11). Suppose, for example, that the disease is emphysema, the exposure is alcohol consumption, and the measure of association is the ratio of emphysema odds among drinkers to that in nondrinkers. Cigarette smoking is a potential confounding factor for the association between emphysema and alcohol, since it is positively correlated with both emphysema risk and alcohol consumption. Confounding of the association by cigarette smoking occurs when the odds-ratio has the same value in smokers and nonsmokers, but a different value in the total population.

In the present setting, an individual's "exposure" is his genotype for the genetic variant of interest, and the potential confounding factor is subpopulation membership. A classic example of such confounding is given by Knowler et al. (1988) in a study of noninsulin-dependent diabetes mellitus (NIDDM) and immunoglobulin haplotypes among residents of the Gila River Indian Community. The subpopulations of this population comprise individuals with varying degrees of Caucasian (vs. Amerindian) heritage. The authors found an inverse association between NIDDM risk and the haplotype $Gm^{3,5,13,14}$. However this haplotype was more prevalent among the largely Causasian subpopulations than among the largely Amerindian ones. Moreover NIDDM prevalence was lower in the Caucasian than the Amerindian subpopulations. When Knowler et al. (1988) adjusted for heritage, the inverse association disappeared.

The hypothetical population of Figure 1 is another example in which the association between disease and a polymorphism is confounded by population stratification. Within each of the two subpopulations, the polymorphism is unassociated with the disease, since disease prevalence is the same among individuals with all three genotypes. In the total population, however, disease prevalence is positively associated with carrier status of allele A. To verify the total population disease prevalences in Figure 1, note that the overall prevalence of disease among AA homozygotes is

$$\Pr(D|AA) = \Pr(pop\ 1|AA)\,(.06) + \Pr(pop\ 2|AA)(.18).$$

Here the proportion of AA homozygotes who belong to population 1 is

$$\Pr(pop\ 1|AA) = \frac{(.4)\,(.01)}{(.4)\,(.01) + (.6)\,(.16)} = .04.$$

Thus disease prevalence in homozygotes is

$$\Pr(D|AA) = (.04)\,(.06) + \Pr(.96)(.18) = .18.$$

Similar calculations show that the prevalences of disease among Aa heterozygotes and aa homozygotes are .16 and .11, respectively. So although there is no association between genotype and disease in each subpopulation, we

Table 1: Distributions of Diseased and Disease-free Individuals according to Genotype and Subpopulation, for the Example of Figure 1

Genotype	Diseased (13.2%)			Disease-free (86.8%)		
	Subpop 1	Subpop 2	Total	Subpop 1	Subpop 2	Total
AA	.002	.131	.133	.004	.091	.095
Aa	.033	.393	.426	.078	.272	.350
aa	.147	.294	.441	.351	.204	.555
Total	.182	.818	1.0	.433	.567	1.0

find one in the total population. Notice that this association is spurious, due entirely to differences between the two subpopulations in prevalence of both disease and of genotype. In this example, allele A and disease are both more common in subpopulation 2 than subpopulation 1. This leads to a positive association in the whole population in the absence of one in each subpopulation. Examples also can be constructed in which there is a positive association in each subpopulation but not in the whole population.

The key conditions for such confounding are differences among the subpopulations with respect to both:

- A) frequencies of genotypes for the variant of interest;
- B) disease prevalence

(Clayton and Hills, 1994; Kelsey et al., 1996, p 11; Pritchard and Rosenberg, 1999). Note that condition (2) of the population stratification model ensures that confounding condition (A) is satisfied for some markers. Therefore population stratification can cause confounding for some polymorphisms and some diseases. In contrast, comparison of condition (A) with the cryptic relatedness property (3) shows that condition (A) fails to hold. Thus confounding by population structure does not occur under the model for cryptic relatedness.

4.2 Variance distortion

Consider a linear combination $Z = \sum_{i=1}^{n} c_i g_i$ of genotypes for n individuals in a population, where the c_i are fixed constants. The variance of Z is

$$var(Z) = \sum_{i=1}^{n} c_i^2 var\left(g_i\right) + \sum_{i \neq i'} c_i c_{i'} cov(g_i, g_{i'}). \qquad (8)$$

We have seen that under the population stratification model, the genotypes of any pair of unrelated individuals are marginally independent; thus their covariances are zero, and $var(Z) = \left(\sum_i c_i^2\right) \sigma^2$, where σ^2 is given by (2). Thus correct specification of σ^2 is enough to correctly specify $var(Z)$. Since consistent estimates of σ^2 are available (Sasieni, 1997), $var\left(Z\right)$ can be estimated consistently in the presence of population stratification. In the presence of cryptic relatedness, however, equation (7) shows that the genotypes of any two individuals are positively correlated. Estimates for $var\left(Z\right)$ used to construct standard test statistics are based on the assumption that genotypes are uncorrelated, and thus they are biased in the presence of cryptic relatedness. The difference between the nominal variance in the absence of correlation and the actual one in its presence is referred to as variance distortion.

5 Implications for Case-control Studies

We now consider the consequences of population stratification and cryptic relatedness on the null distributions of test statistics based on case-control data. Because case-control studies sample the genotype distributions of diseased and disease-free individuals in the population, we begin by considering the impact of the two forms of structure on the population mean genotypes of diseased and disease-free individuals.

5.1 Population parameters

Suppose we wish to conduct a case-control study by sampling the population of Figure 1. This would involve sampling the genotype distributions shown in columns 4 and 7 of Table 1. These distributions are $(.133, .426, .441)$ for the diseased and $(.095, .350, .550)$ for the disease-free. (To verify the entries of Table 1, note first from Figure 1 that the diseased group comprises $\left[(.4)\,(.06) + (.6)\,(.18)\right] \times 100 = 13.2\%$ of the population. Thus, based on the subpopulation-specific genotype prevalences in Figure 1, we find that the fraction of diseased individuals in subpopulation 1 who carry genotype AA is $(.4)\,(.01)\,(.06)\,/\,(.132) = .002$. The remaining entries are calculated similarly.)

Table 2 gives notation for the population parameters in a general version of Table 1. In this table, π_{djg} denotes the prevalence of genotype g among diseased ($d = 1$) or disease-free ($d = 0$) individuals in subpopulation j, $j = 1, 2$. (For the data in Table 1, for example, the genotype

Table 2: Distributions of Diseased and Disease-free Individuals according to Genotype and Subpopulation, for a Diallelic Polymorphism and Two Subpopulations

Genotype	Diseased ($d=1$)			Disease-free ($d=0$)		
	Subpop 1	Subpop 2	Total	Subpop 1	Subpop 2	Total
AA	$a_{11}\pi_{112}$	$a_{12}\pi_{122}$	$a_{11}\pi_{112}+a_{12}\pi_{122}$	$a_{01}\pi_{012}$	$a_{02}\pi_{022}$	$a_{01}\pi_{012}+a_{02}\pi_{022}$
Aa	$a_{11}\pi_{111}$	$a_{12}\pi_{121}$	$a_{11}\pi_{111}+a_{12}\pi_{121}$	$a_{01}\pi_{011}$	$a_{02}\pi_{021}$	$a_{01}\pi_{011}+a_{02}\pi_{021}$
aa	$a_{11}\pi_{110}$	$a_{12}\pi_{120}$	$a_{11}\pi_{110}+a_{12}\pi_{120}$	$a_{01}\pi_{010}$	$a_{02}\pi_{020}$	$a_{01}\pi_{010}+a_{02}\pi_{020}$
Total	a_{11}	a_{12}	1	a_{01}	a_{02}	1

prevalences in subpopulation 1 are $(\pi_{d10}, \pi_{d11}, \pi_{d12}) = (.01, .18, .81)$, independent of disease status d.)

To compare the genotype prevalences π_{1jg} and π_{0jg} in the two disease groups, we need a model for the relation between genotype and disease risk. We shall assume an additive logistic model that gives disease risk, conditional on genotype and subpopulation membership, as

$$\text{logit } P(d=1|g, S=j) = \alpha_j + \beta g, g = 0, 1, 2,\ j = 1, 2. \tag{9}$$

This model specifies that the heterozygote odds-ratio e^{β} is constant over all subpopulations. Absence of association (the null hypothesis) occurs when $\beta = 0$. Although β often is used to quantify the magnitude of the association, it will be more convenient to quantify it here by the difference δ in mean genotypes between diseased and disease-free individuals. When $\beta = 0$, diseased and disease-free individuals in a specific subpopulation j have the same genotype frequencies $\pi_{1jg} = \pi_{0jg}$, $g = 0, 1, 2$. Thus, conditional on subpopulation, they have the same mean genotypes $\mu_{dj} = 2\pi_{dj2} + \pi_{dj1}$, and so the difference $\mu_{1j} - \mu_{0j}$ in mean genotypes between diseased and disease-free groups is zero. In the presence of confounding, however, the two disease groups may have different marginal mean genotypes in the entire population. For example, the prevalences in Table 1 imply that, within each of the two subpopulations, the mean genotype is independent of disease status (the common mean for individuals in subpopulation 1 is $[2(.002) + .033] / (.182) = [2(.004) + .078] / (.433) = .20$, and the corresponding common mean for those in subpopulation 2 is .80). In the overall population, however, diseased and disease-free groups have mean genotypes $2(.133) + .426 = .692$ and $2(.095) + .350 = .540$, respectively, giving a difference of $\delta = .692 - .540 = .152$. Conversely, examples can be constructed in which diseased and disease-free groups have unequal mean genotypes within each subpopulation ($\beta \neq 0$), but equal ones in the population as a whole.

Let $\delta_j = \mu_{1j} - \mu_{0j}$ denote the difference in mean genotypes between diseased and disease-free individuals in subpopulation j, $j = 1, 2$. The overall difference in mean genotypes between the two groups can be written

$$\delta = \sum_j a_{1j}\mu_{1j} - \sum_j a_{0j}\mu_{0j} = \overline{\delta} + \Delta. \tag{10}$$

Here

$$\overline{\delta} = a_{11}\delta_1 + a_{12}\delta_2 \tag{11}$$

denotes the average of the subpopulation-specific differences, weighted by the proportions a_{1j} of diseased individuals in the subpopulations. Also the bias term

$$\Delta = (\mu_{02} - \mu_{01})(a_{01} - a_{11}) \tag{12}$$

is the product of two factors: 1) the difference $\mu_{02} - \mu_{01}$ in mean genotypes of disease-free individuals in subpopulations 2 and 1; and 2) the difference $a_{01} - a_{11}$ in subpopulation 1 membership between disease-free and diseased individuals. Confounding occurs when $\Delta \neq 0$. In this case, confounding can cause rejection of the null hypothesis even when the average difference $\overline{\delta} = 0$. (This is the case for the example of Figure 1 and Table 1, where $\overline{\delta} = 0$ and $\Delta = (.368)(.257) = .092$). Moreover if $\overline{\delta}$ and Δ are both nonzero and have opposite sign, confounding can cause loss of power.

We can now restate the conclusions of the preceding section in terms of the genotypes of diseased and disease-free individuals. Equation (12) shows that a necessary condition for confounding is that $\mu_{01} \neq \mu_{02}$ (mean genotypes in the disease-free group differ within the two subpopulations). For rare diseases, confounding condition (A), which stipulates different subpopulation genotype frequencies overall, implies different frequencies among the disease-free, which is necessary for a difference $\mu_{01} \neq \mu_{02}$ in their mean genotypes within subpopulations. The population stratification model satisfies condition (A) and therefore this form of structure poses a potential confounding problem. In contrast, the cryptic relatedness model fails to satisfy condition (A), and hence this form of structure cannot cause confounding.

Equation (12) also gives a second requirement for confounding, namely that $a_{11} \neq a_{01}$ (diseased and disease-free individuals are distributed differently within the two subpopulations). Stated differently, this requirement states that the two subpopulations must differ in disease prevalence, which is confounding condition (B).

5.2 The distributions of test statistics

We now explore how population stratification and cryptic relatedness affect the distribution of a commonly used test statistic for case-control data, the Armitage trend statistic (Armitage, 1955). Suppose that N diseased cases and N disease-free controls have been genotyped for a polymorphism of interest. Let G_{di} denote the genotype of the i^{th} individual in group d, $d = 0, 1$, $i = 1, \ldots, N$, and let

$$Z = N^{-1/2} \left(\sum_{i=1}^{N} G_{1i} - \sum_{i=1}^{N} G_{0i} \right)$$

denote the (scaled) mean difference in genotype counts between cases and controls. The (squared) Armitage trend statistic for testing the null hypothesis of no disease-genotype association is

$$X^2 = \frac{Z^2}{2\widehat{\sigma}^2}, \tag{13}$$

where

$$\widehat{\sigma}^2 = \frac{1}{2N} \sum_{d=0}^{1} \sum_{i=1}^{N} G_{di}^2 - \left(\frac{1}{2N} \sum_{d=0}^{1} \sum_{i=1}^{N} G_{di} \right)^2. \tag{14}$$

(Armitage, 1955; Devlin and Roeder, 1999).Significance levels are obtained by referring X^2 to a central chi-square distribution on one degree of freedom: $X^2 \sim \chi_1^2(0)$.

In the presence of population stratification, Z has marginal mean

$$E[Z] = N^{1/2} \left(\overline{\delta} + \Delta \right),$$

where $\overline{\delta}$ and Δ are given by equation (10). We have seen in Section 2 that for this model genotypes of different individuals are uncorrelated. Thus the variance of Z is

$$v_1 = 2var(G) = 2\sigma^2,$$

where σ^2 is given by equations (2) and (3). Since $\widehat{\sigma}^2$ of (14) is consistent for σ^2 under the null hypothesis $\overline{\delta} = 0$ and in the absence of confounding ($\Delta = 0$), the trend statistic (13) is distributed asymptotically as $X^2 \sim \chi_1^2(0)$ in this circumstance. In general, the distribution of X^2 is approximately $\chi_1^2(\phi)$, with noncentrality parameter

$$\phi = \frac{N \left(\overline{\delta} + \Delta \right)^2}{2\sigma^2}. \tag{15}$$

Expression (15) indicates that for large sample sizes, even a small amount of confounding, as reflected in the magnitude of Δ, could lead to appreciable inflation of the type-1 error rate when $\overline{\delta} = 0$, or loss of power when $\overline{\delta} \neq 0$ and $\overline{\delta}$ and Δ have opposite sign.

In the presence of cryptic relatedness, Z has marginal mean equal to $N^{1/2}\overline{\delta}$. As shown by Devlin and Roeder (1999), the null marginal variance of Z is

$$\begin{aligned} Nvar(Z) &= \sum_{i=1}^{N} [var(G_{1i}) + var(G_{0i})] \\ &\quad + \sum_{1 \le i < i' \le N} [cov(G_{1i}, G_{1i'}) + cov(G_{0i}, G_{0i'})] \\ &\quad - 2 \sum_{i=1}^{N} \sum_{i'=1}^{N} cov(G_{1i}, G_{0i'}). \end{aligned}$$

Equations (6) and (7) show that $var(G_{di}) = 2pq(1+f)$, and $cov(G_{di}, G_{di'}) = \sum_{j=1}^{2} a_{dj}^2 (4fpq)$, $i \neq i'$, $d = 0, 1$. Also, $cov(G_{1i}, G_{0i'}) = \sum_{j=1}^{2} a_{0j} a_{1j} (4fpq)$. Thus for large N, $var(Z)$ is approximately,

$$v_2 = 4pq(1+f) + 8fNpq(a_{11} - a_{01})^2.$$

Under the null hypothesis $\overline{\delta} = 0$, the asymptotic distribution of X^2 has the mean of a $\chi_1^2(0)$ variable but not its variance. As noted by Devlin and Roeder (1999)the variable $\lambda^{-1}X^2$ has a $\chi_1^2(0)$ distribution, where

$$\lambda = \frac{v_2}{v_1} = 1 + \frac{2fN}{1+f}(a_{11} - a_{01})^2$$

is the ratio of variances under the cryptic relatedness and stratification models. To correct the trend statistic X^2 for possible biased estimation of v_2, the GC method estimates λ using genotypes at other unlinked markers, and then scales X^2 as $\widehat{\lambda}^{-1}X^2$. Thus the GC method corrects for potential variance misspecification due to cryptic relatedness, but not for potential confounding bias due to population stratification.

This distinction is illustrated in Figure 2, which shows the distribution of the trend statistic X^2 for 1000 cases and 1000 controls in the presence and absence of an association between the disease and the variant, and in the presence and absence of confounding due to stratification into two subpopulations of equal size. The disease risk in the whole population is fixed at $K = 5\%$, and normal homozygotes in subpopulation 2 are assumed to have three times the risk of those in subpopulation 1. In all panels of the figure the true value of λ is one, so that the GC-adjusted statistic is asymptotically equivalent to the trend statistic.

Panel A shows the distribution of X^2 in the absence of both association and confounding. In this case X^2 has a central $\chi_1^2(0)$ distribution, and so the probability that X^2 exceeds the critical value 3.84 (or 10.83) corresponding to a type-1 error rate $\alpha = .05$ (or $\alpha = .001$) has its correct value α.

Panel B shows the distribution of X^2 in the absence of causal association but in the presence of confounding. Also shown are the actual type-1 error rates corresponding to nominal rates of $\alpha = .05$ and $\alpha = .001$. Now X^2 has a noncentral chi-squared distribution $\chi_1^2(\phi)$, with

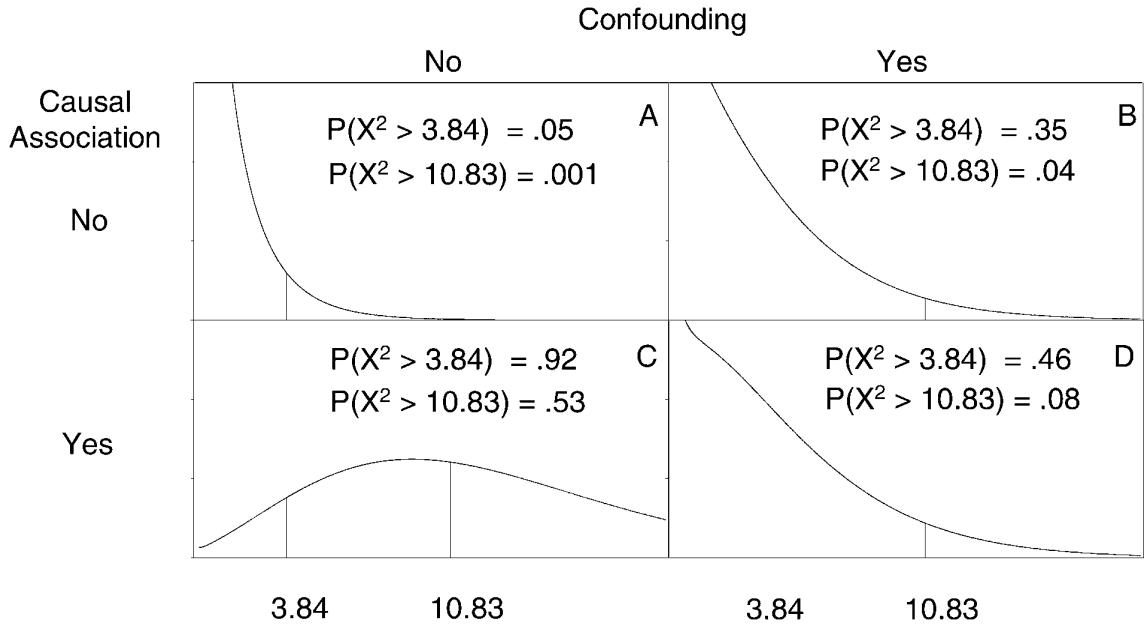

Figure 2: Actual distribution of trend statistic X^2 for 1000 cases and 1000 controls, with and without a true causal association and with and without confounding by population stratification. Graphs show chi-square distributions on one degree of freedom with noncentrality parameters given by: A) 0; B) 2.47; C) 11.30; and D) 3.48.

noncentrality parameter $\phi = N\Delta^2/\left(2\sigma^2\right)$. The magnitude and sign of the confounding parameter Δ are determined by the difference in variant allele frequencies in the two subpopulations, as described in the Appendix. The value $\phi = 2.47$ in Panel B results when the variant has frequency 5% and 10% in subpopulations 1 and 2, respectively. Comparison of panels A and B shows that confounding can cause substantial inflation of the type-1 error rate α, and that the inflation factor increases as α decreases.

Panel C of Figure 2 shows the chi-squared distribution of X^2 in the presence of a causal association and in the absence of confounding. Here the noncentrality parameter is $\phi = N\overline{\delta}^2/\left(2\sigma^2\right)$. Its value $\phi = 11.30$ corresponds to a common variant frequency of 5% and a common heterozygote relative risk of 1.5 in the two subpopulations (see Appendix). Panel C also shows the power of the trend test with type-1 error rates $\alpha = .05$ and $\alpha = .001$. The power is 92% for a test of size $\alpha = .05$ and 43% for a test of size $\alpha = .001$.

Panel D shows the distribution of X^2 in the presence of both confounding and association, with a heterozygote relative risk of 1.5 in each of the subpopulations. Here the noncentrality parameter is $\phi = N\left(\overline{\delta} + \Delta\right)^2/\left(2\sigma^2\right)$. Its value $\phi = 3.48$ is lower than the one for Panel C because $\overline{\delta}$ and Δ have opposite sign. This occurs when, compared to subpopulation 1, subpopulation 2 has higher disease prevalence but lower frequency of the variant. (Here the variant has frequency 10% in subpopulation 1 but only 5% in subpopulation 2.) Comparison of panels C and D shows that such "negative" confounding can cause appreciable power loss, with the percentage loss increasing as the type-1 error rate becomes more stringent.

These examples support the assertions of Marchini, et al. (2004) and Helgason et al., (2005) that confounding should not be dismissed in case-control studies using large sample sizes to detect small effects. Since the examples were constructed for the circumstance that the trend statistic and the GC-adjusted statistic are asymptotically equivalent, they also show that GC adjustment methods do not correct for confounding bias due to population stratification.

6 Implications for Family Studies

Study designs that compare phenotypes of diseased individuals to those of their unaffected siblings or to untransmitted parental genotypes are matched with respect to subpopulation membership. Therefore these designs avoid confounding bias due to population stratification. However discordant sibling designs are vulnerable to variance distortion arising from cryptic relatedness. This distortion can occur because discordant sibling test statistics are based on the assumption that genotypes of individuals in distinct sibships are uncorrelated, in violation of the positive correlation specified by the cryptic relatedness model. In contrast, the transmission disequilibrium test (TDT) statistic avoids both confounding bias and variance distortion by conditioning on observed parental genotypes or by other statistical methods (Rabinowitz and Laird, 2000; Rabinowitz, 2002; Whittemore and Halpern, 2003).

7 Guidelines for Case-Control Analysis

Cryptic relatedness may be appreciable in populations with extensive nonrandom mating, such as diseased individuals who inherit a common mutation, or small endogamously mating populations. (In fact, variance distortion due to cryptic relatedness can occur even without subdivision of the population into subpopulations.) While cryptic relatedness could, in principle, be a serious source of incorrect inference, recent work suggests that this is unlikely to be the case for outbred populations. Specifically, since allelic correlation reflects identity-by-descent of chromosomal segments, the magnitude of such correlation in two randomly selected chromosomes from a given population depends on the average number of prior generations leading back to a common ancestor. Voight and Pritchard (2005) proposed a coalescent model for cryptic relatedness, and estimated the model parameters in both inbred and outbred populations. They concluded that while cryptic relatedness may be troublesome for inbred populations, it is less important in outbred ones. Accordingly, this section focuses on how to deal with confounding bias due to population stratification.

The standard way to control confounding in case-control studies is to construct an adjusted test statistic by matching or stratifying on levels of the confounding variable, or by modelling the joint dependence of disease status on exposure and confounder (Breslow and Day, 1980). However these methods require knowledge about the subpopulation to which study subjects belong. Several investigators have proposed latent variable methods for inferring subpopulation membership, so that standard methods can be used (e.g. Pritchard et al. (1999, 2001), Satten et al. (2001), Schork et al. (2001)). Satten et al. (2001) use a likelihood-based method and the EM algorithm, treating the population membership of study subjects as missing data. These authors use simulated data to evaluate the efficacy of this approach. They note that the large number of parameters involved creates difficulties in finding global maxima for the likelihood function, and recommend a principal components approach to choosing good initial values for the algorithm. Recently Wang et al. (2005) proposed a stratified analysis conditional on genotypes of a single null marker unlinked to the locus of interest, effectively using these genotypes as surrogates for subpopulation membership. All the latent variable methods are vulnerable to misspecification of subpopulation membership and consequent residual confounding. They also involve unverifiable assumptions, and most involve many unknown parameters. There is need for further evaluation of their performance with data that are confounded by population structure. The few available results suggest that some of them may not do well (Marchini et al., 2004; Campbell et al., 2005; Pritchard and Donnelly, 2001).

A strategy that avoids the use of adjusted test statistics would instead use a sensitivity analysis based on crude assumptions about the extent of confounding plausible in a given study (Wacholder et al., 2000). For example, the Appendix provides a formula for the noncentrality parameter ϕ of (15) for the null distribution of the Armitage trend statistic, in the presence of confounding due to stratification into two subpopulations of equal size. ϕ is given in terms of the number N of cases and controls, the allele frequencies p_1 and p_2 of the deleterious allele in subpopulations 1 and 2, the ratio ρ of disease risk in subpopulation 2 relative to subpopulation 1, and the population disease prevalence K. For a given combination of these parameters and a given type-1 error level, one can compute the corresponding critical value of the actual null distribution, and reject the null hypothesis only if the test statistic exceeds this value. Table 3 shows such critical values for a range of parameter values. Tabulations such as this could be used to determine the level of confounding needed to cast doubt on a nominally significant finding, or analogously, to argue that an observed test statistic exceeds the critical value corresponding to the maximum plausible level of stratification. In principle, sample size needs could be targeted to give adequate power for detecting a case-control genotype difference large enough to rule out confounding of plausible magnitudes.

Recently, Gorroochurn, et al. (2006) proposed estimating the confounding parameter Δ for a candidate marker by using a set of L unlinked diallelic markers. The estimate $\widehat{\Delta}$, which is proportional to the mean case-control difference in minor allele frequencies at the L markers, is used to "centralize" the chi-squared distribution of the test statistic at the candidate marker. The utility of this strategy involves an assumption most easily described for a population stratified into $J = 2$ subpopulations. In this case, Δ is proportional to the inter-subpopulation difference $d = p_2 - p_1$ in variant allele frequency at the candidate marker (see Appendix equation (26)). In contrast, $\widehat{\Delta}$ is proportional to the mean inter-subpopulation difference $\overline{d} = L^{-1} \sum_{\ell=1}^{L} d_\ell$ in minor allele frequencies at the L markers. Using $\widehat{\Delta}$ to estimate Δ involves the questionable assumption that $\overline{d} = d$.

8 Discussion

The preceding arguments have shown that, loosely speaking, population stratification can distort the mean of the Armitage trend statistic but not its variance, while cryptic relatedness can distort its variance but not its mean. This difference has two consequences. First, the methods proposed for dealing with cryptic relatedness do not adjust for confounding bias due to population stratification. Indeed, as noted above, this bias cannot be addressed reliably by using a panel of markers unlinked to the locus of interest, a strategy that underlies these methods. This is because bias depends on inter-subpopulation differences in allele frequencies specifically at the marker of interest, which cannot reliably be estimated from differences at other markers. This local specificity has been

Table 3: Critical Values of Actual Null Distribution of the Armitage Trend Statistic[a] with Stratification into Two Subpopulations of Equal Size

			Significance Level		
			.05	.01	.001
No Confounding			3.84	6.63	10.83
Confounding					
deleterious allele frequency		ratio of disease prevalence in			
subpop 1	subpop 2	subpop 1 to that in subpop 2[b]			
.01	.02	2	4.61	7.86	12.64
		3	5.48	9.09	14.25
	.03	2	5.96	9.73	15.06
		3	8.01	12.34	18.28
.05	.10	2	7.26	11.39	17.13
		3	10.35	15.20	21.74
	.15	2	12.03	17.23	24.15
		3	19.19	25.63	33.95
.10	.20	2	10.14	14.94	21.43
		3	15.63	21.48	29.14
	.30	2	18.80	25.17	33.42
		3	32.28	40.49	50.79
.15	.30	2	13.01	18.39	25.52
		3	21.04	27.76	36.39
	.45	2	25.98	33.39	42.80
		3	46.55	56.31	68.36

a) based on 1000 cases and 1000 controls
b) assuming an overall disease prevalence of .05

demonstrated in a population of European-Americans (Campbell et al., 2005). Second, family-based designs such as discordant sib pairs, which match cases to controls with respect to subpopulation membership, control for confounding by population stratification, but they do not adjust for variance distortion due to cryptic relatedness.

The models for population stratification and cryptic relatedness are both simplified approximations to reality and neither is apt to fully describe the population structure relevant to a particular population. It is possible to construct more general models that combine the features of the two models and to develop test statistics robust against both types of population structure (see for example Devlin et al. (2001)). However this robustness would be gained at the price of additional analytic complexity and potential power loss.

An alternative strategy would be to evaluate the level of confounding needed to vitiate a positive finding, and determine whether that level is plausible in the given population (see Wacholder et al. (2000) for further discussion). This strategy involves the type of crude sensitivity analysis illustrated in Table 3; further refinement of the assumptions may be helpful. However, while the strategy may provide some reassurance about positive findings, it does not address the problem of false negatives due to confounding, a source of potentially serious power loss.

In summary, a major issue for the future of genetic association studies is the extent to which subpopulations of the major racial/ethnic groups differ in both disease prevalence and allele frequencies. At present little data are available to address this question (see Helgason et al. (2005) and Campbell et al. (2005) for exceptions). These studies indicate that population stratification can exist in Caucasian populations, and that it can cause confounding. Thus, further characterization of population structure in major racial/ethnic groups and evaluation of methods for dealing with it should be a major priority for the next generation of efforts to characterize human genetic variation.

APPENDIX: The distribution of the Armitage trend statistic in the presence of population stratification

We describe the distribution of the Armitage trend statistic when applied to the genotypes of N cases and N controls sampled from a population stratified into two subpopulations of equal size, without cryptic relatedness. We have seen in Section 5.2 that in this case the statistic

has asymptotically a noncentral chi-squared distribution on one degree of freedom. Its noncentrality parameter

$$\phi = \frac{N\left(\overline{\delta} + \Delta\right)^2}{2\sigma^2} \tag{16}$$

determines the shapes of the curves in Figure 2, and the critical values shown in Table 3. The value of ϕ depends on the level of association between variant and disease risk (measured by the parameter $\overline{\delta}$), the level of confounding (measured by the parameter Δ) and the genotype variance σ^2. Our objective is to express the quantities $\overline{\delta}$, Δ and σ^2 in terms of the Hardy-Weinberg (HW) variant allele frequencies p_1 and p_2 in the two subpopulations, the disease prevalence K in the whole population, the risk ratio ρ among normal homozygotes in subpopulation 2 relative to risk among those in subpopulation 1, and the relative risk γ among heterozygote carriers of the variant compared to normal homozygotes.

We assume the additive logistic model (9) for the relation between disease risk and genotype within each subpopulation. We also assume a small disease risk in each subpopulation, so that an individual from subpopulation j with genotype g has approximate risk

$$r_j \gamma^g, \quad \text{where} \quad r_j = e^{\alpha_j} \text{ and } \quad \gamma = e^{\beta}. \tag{17}$$

Under the multiplicative model (17), the disease risk in subpopulation j can be written

$$\begin{aligned} \Pr\left(D = 1 | \text{ subpop } j\right) &= r_j \left[p_j^2 \gamma^2 + 2 p_j q_j \gamma + q_j^2\right] \\ &= r_j \left[1 + p_j \left(\gamma - 1\right)\right]^2 \equiv r_j \eta_j. \end{aligned} \tag{18}$$

Hence the overall disease risk is

$$K = .5\left(r_1 \eta_1 + r_2 \eta_2\right) = .5 r_1 \eta_1 \left(1 + \rho\nu\right),$$

where $\rho = r_2 / r_1$ and $\nu = \eta_2 / \eta_1$.We can thus write the probabilities that a diseased individual belongs to each of the subpopulations as

$$a_{11} = \frac{.5 r_1 \eta_1}{K} = \frac{1}{1 + \rho\nu} \text{ and } a_{12} = 1 - a_{11} = \frac{\rho\nu}{1 + \rho\nu}. \tag{19}$$

Also, the multiplicative model (17) implies that the variant allele has HW frequency $p_j \gamma / \left[1 + p_j \left(\gamma - 1\right)\right]$ among diseased individuals in subpopulation j, $j = 1, 2$ (Clayton, 1999). Hence the mean genotype among diseased cases in subpopulation j is

$$\mu_{1j} = \frac{2 p_j \gamma}{1 + p_j \left(\gamma - 1\right)}.$$

The mean genotype among disease-free individuals in the subpopulation is approximately $\mu_{0j} = 2 p_j$, giving a difference

$$\delta_j = \mu_{1j} - \mu_{0j} = \frac{2 p_j q_j \left(\gamma - 1\right)}{1 + p_j \left(\gamma - 1\right)}. \tag{20}$$

Combining (11), (19) and (20), we have

$$\begin{aligned} \overline{\delta} &= a_{11}\delta_1 + a_{12}\delta_2 = \frac{\delta_1 + \rho\nu\delta_2}{1 + \rho\nu} \\ &= \frac{2\left(\gamma - 1\right)}{1 + \rho\nu}\left[\frac{p_1 q_1}{1 + p_1\left(\gamma - 1\right)} + \rho\nu \frac{p_2 q_2}{1 + p_2\left(\gamma - 1\right)}\right]. \end{aligned} \tag{21}$$

Next we express the confounding parameter Δ of (12) in terms of the variant allele frequencies and disease risks. Because disease prevalence K is low, the difference $\mu_{02} - \mu_{01}$ in mean genotypes among disease-free individuals in subpopulations 2 and 1 is approximately $2\left(p_2 - p_1\right)$. Substituting this expression into equation (12) gives

$$\Delta = 2\left(p_2 - p_1\right)\left(a_{01} - a_{11}\right), \tag{22}$$

where a_{11} and a_{01} are the fractions of a diseased and disease-free individuals, respectively who belong to subpopulation 1. Arguments analogous to those preceding equations (19) give the fraction of disease-free individuals who belong to subpopulation 1 as

$$a_{01} = \frac{.5\left[1 - \frac{2K}{1 + \rho\nu}\right]}{1 - K}.$$

Thus, after some algebra, we find that

$$a_{01} - a_{11} = \frac{.5\left(\rho\nu - 1\right)}{\left(1 + \rho\nu\right)\left(1 - K\right)}. \tag{23}$$

Substituting (23) into (22) gives

$$\Delta = \frac{\left(p_2 - p_1\right)\left(\rho\nu - 1\right)}{\left(1 + \rho\nu\right)\left(1 - K\right)}. \tag{24}$$

Finally, we approximate the variance term in the denominator of ϕ by its value under the null hypothesis of no association. Using $a_1 = a_2 = 0.5$ in equations (2) and (3) gives

$$\sigma^2 = p_1 + p_2 - 2 p_1 p_2. \tag{25}$$

In conclusion, substituting expression (21) for $\overline{\delta}$, expression (24) for Δ and expression (25) for σ^2 into (16) completes the calculation of ϕ.

Under the null hypothesis $\gamma = 1$ of no association between variant and disease, the expressions for $\overline{\delta}$ and Δ simplify to $\overline{\delta} = 0$ and

$$\Delta = \frac{\left(p_2 - p_1\right)\left(\rho - 1\right)}{\left(1 + \rho\right)\left(1 - K\right)}. \tag{26}$$

ACKNOWLEDGEMENTS

This work was supported by NIH grant number CA94069. The author is grateful to Raymond R. Balise for technical support and to Joseph B. Keller for helpful discussions.

REFERENCES

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

Breslow, N.E. and Day, N.E. (1980), "Statistical methods in cancer research, Volume I - The analysis of case-control studies," *IARC Scientific Publications* (32):5-338.

Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G., Hirschhorn, J.N. (2005), "Demonstrating stratification in a European American population," *Nature Genetics*, **37**, 868-872.

Clayton, D. (1999), "A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission," *American Journal of Human Genetics*, **65**, 1170-1177.

Clayton, D. and Hills, M. (1994), *Statistical Models in Epidemiology*, Oxford Science Publications.

Crow, J.F. and Kimura, H. (1970), *An Introduction to Population Genetics Theory*, Burgess Publishing Co., Minneapolis, Minnesota.

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

Devlin, B., Roeder, K., Wasserman, L. (2001), "Genomic control, a new approach to genetic association studies," *Theoretical Population Biology*, **60**, 156-168.

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

Gorroochurn, P., Heiman, G.A., Hodge, S.E., Greenberg, D.A. (2006), "Centralizing the non-central chi-square: A new method to correct for population stratification in genetic association studies," *Genetic Epidemiology*, **30**, 277-289.

Helgason, A., Yngvadottir, B., Hrafnkelsson, B., Gulcher, J., Stefansson, K. (2005), "An Icelandic example of the impact of population structure on association studies," *Nature Genetics*, **37**, 90-95.

Kelsey, J.L., Whittemore, A.S., Evans, A.S., Thompson, W.D. (1996), *Methods in Observational Epidemiology*, Second Edition, Oxford University Press, New York.

Knowler, W.C., Williams, R.C., Pettitt, D.J., Steinberg, A.G. (1988), "GM3;5, 13, 14 and type 2 diabetes mellitus: An association in American Indians with genetic admixture," *American Journal of Human Genetics*, **43**, 520-526.

Marchini, J., Cardon, L.R., Phillips, M.S., Donnelly, P. (2004), "The effects of human population structure on large genetic association studies," *Nature Genetics*, **36**, 512-517.

Pritchard, J.K. and Rosenberg, N.A. (1999), "Use of unlinked genetic markers to detect population stratification in association studies," *American Journal of Human Genetics*, **65**, 220-228.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P. (2001), "Association mapping in structured populations," *American Journal of Human Genetics*, **67**, 170-181.

Pritchard, J.K. and Donnelly, P. (2001), "Case-control studies of association in structured or admixed populations," *Theoretical Population Biology*, **60**, 227-237.

Rabinowitz, D. (2002), "Adjusting for population heterogeneity and misspecified haplotype frequencies when testing nonparametric null hypotheses in statistical genetics," *Journal of the American Statistical Association*, **97**, 742-758.

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," *Human Heredity*, **50**, 211-223.

Reich, D.E. and Goldstein, D.B. (2001), "Detecting association in a case-control study while correcting for population stratification," *Genetic Epidemiology*, **20**, 4-16.

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

Satten, G.A., Flanders, W.D., Yang, Q. (2001), "Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model," *American Journal of Human Genetics*, **68**, 466-477.

Schork, N.J., Fallin, D., Thiel, B., Xu, X., Broeckel, U., Jacob, H.F., Cohen, D. (2001), "The future of genetic case-control studies" (Review), *Advances in Genetics*, **42**, 191-212.

Setakis, E., Stirnadel, H., Balding, D.J. (2006), "Logistic regression protects against population structure in genetic association studies," *Genome Research*, **16**, 290-296.

Voight, B.F., Pritchard, J.K. (2005), "Confounding from cryptic relatedness in case-control association studies," *PLoS Genetics,* **1**, e32.

Wacholder, S., Rothman, N., Caporaso, N. (2000), "Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias," *Journal of the National Cancer Institute*, **92**, 1151-1158.

Wang, Y., Localio, R., Rebbeck, T.R. (2005), "Bias correction with a single null marker for population stratification in candidate gene association studies," *Human Heredity*, **59**, 165-175.

Whittemore, A.S. and Halpern, J. (2003), "Genetic association tests for family data with missing parental genotypes: A comparison," *Genetic Epidemiology*, **25**, 80-91.

Zheng, G., Freidlin, B., Gastwirth, J.L. (2006), "Robust genomic control for association studies," *American Journal of Human Genetics*, **78**, 350-356.