

Point/CounterpointPoint: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations?<sup>1</sup>Duncan C. Thomas<sup>2</sup> and John S. Witte

Department of Preventive Medicine, University of Southern California, Los Angeles, California 90033-9987 [D. C. T.], and Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio 44106-4945 [J. S. W.]

**Introduction**

The case-control design is a widely used approach for investigating associations between candidate genes and dichotomous disease traits. As most commonly implemented, cases are drawn from a population-based disease registry and unrelated controls from their source population, perhaps matching on age, gender, ethnicity, or other potential confounders. As has been pointed out by various authors (1–6), such designs are susceptible to a form of confounding known in the genetics literature as “population stratification” if the gene under study shows marked variation in allele frequency across subgroups of the population (at least within levels of the matching factor) and if these subgroups also differ in their baseline risk of the disease. Several extreme examples of such confounding have been widely discussed in the literature and are revisited briefly below. At the present time, however, little is known about the extent and implications of this phenomenon in less extreme situations. In this report and the “Counterpoint” that follows (7), we discuss the potential seriousness of this concern and suggest approaches to dealing with it.

Before proceeding, we wish to set the context of this contribution. Gene association studies (whether case-control or cohort) can be used either in an “indirect” manner as a tool for mapping genes using linkage disequilibrium or in a “direct” manner for evaluating associations with postulated causal (“candidate”) genes.<sup>3</sup> Both indirect and direct associations are subject to the same potential bias attributable to population stratification,<sup>4</sup> but here we are primarily concerned with the latter type of study. False-positive associations with markers that are in linkage disequilibrium with a causal gene, as can

easily arise in recently admixed populations for example, will often not be replicated in different populations but are nevertheless “interesting” as an indication that a causal gene may be in the general region. [In fact, studying admixed populations can benefit linkage disequilibrium mapping (8).] Although some effort by the scientific community may be wasted trying to replicate such reports, this may still be rewarding in the context of the larger gene mapping effort in terms of further localization of such genes. On the other hand, false-positive associations with candidate genes are essentially dead ends, and if a high proportion of such associations turn out to be false positives, the wasted effort could be considerable.

**Population Stratification: The Potential Problem**

**Classic Examples.** Several candidate gene association studies are commonly cited as examples of population stratification. One classic example is given by Knowler *et al.* (9), who showed that a failure to adjust for confounding by population stratification would produce a spurious inverse association between variants in the immunoglobulin haplotype  $Gm^{3,5,13,14}$  and non-insulin-dependent diabetes mellitus among residents of the Gila River Indian Community. This association was not causal and instead reflected confounding by a population-stratifying factor, degree of Caucasian heritage. In particular, the inverse association actually reflected the association between heritage and  $Gm^{3,5,13,14}$  and the inverse association between Caucasian heritage and risk of non-insulin-dependent diabetes mellitus. When Knowler *et al.* (9) adjusted for heritage, the inverse association disappeared.

As another example, numerous studies of the purported association between the  $A1$  allele at the  $D_2$  dopamine receptor locus ( $DRD2$ ) and alcoholism give equivocal results; initial reports strongly suggested an association, whereas further studies failed to support this finding. Gelernter *et al.* (10) evaluated published studies attempting to replicate the initial association observed by Blum *et al.* (11) and found much greater heterogeneity among studies than differences between alcoholics and controls. This result might be explained by population stratification, because there are large ethnic differences in  $DRD2$  alleles, from 10% among Yemenite Jews to 80% among Cheyenne Indians; among the controls in the 11 studies reviewed by Gelernter *et al.* (10), the frequencies ranged from 6 to 24% (10 to 37% among cases). Moreover, there is a wide range of ethnic differences in the incidence of alcoholism. The few studies that were restricted to ethnically homogeneous populations did not observe an association (11). Two recent reports, both using family-based study designs, also found no association with  $DRD2$  alleles (12, 13) and provide a comprehensive discussion of ethnic variation in allele frequencies and literature on this association. Although these examples imply that population stratification might be a serious concern in genetic association studies, the potential magnitude of the bias resulting from this phenomenon remains unclear.

A somewhat different perspective on population stratifi-

Received 3/12/01; revised 2/27/02; accepted 3/18/02.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

<sup>1</sup> Supported in part by NIH Grants R01-CA52862, P30-ES07048, R01-CA88164, and R29-CA73270.

<sup>2</sup> To whom requests for reprints should be addressed, at Department of Preventive Medicine, University of Southern California, 1540 Alcazar Street, CHP-220, Los Angeles, CA 90033-9987.

<sup>3</sup> By “causal,” we mean that all other things being equal, including environmental factors and other genes, an individual’s disease risk depends upon his genotype at that locus.

<sup>4</sup> By “bias,” we mean a deviation of the expected value of a parameter estimated from repeated samples using the same study design from its true population value, the parameter of interest here being the genetic relative risk. It can also be used to refer to a deviation of the rate of rejection of the null hypothesis from the test’s nominal value  $\alpha$ . By “population stratification,” we mean the subdivision of a study population into unidentified subgroups with differing allele frequencies and baseline risks of disease.

cation is given by a third classic example. An association has been reported between the 5'FP RFLP adjacent to the insulin gene on chromosome 11p and insulin-dependent diabetes mellitus (14). Although this association had been consistently replicated across several populations, standard affected sib pair methods initially gave no evidence of linkage (15). This situation suggested that the association results might be attributable to population stratification and led Spielman *et al.* (16) to implement the TDT,<sup>5</sup> which showed highly significant evidence of linkage disequilibrium. It appears that the initial failure of the affected sib pair linkage tests was attributable to their low power for detecting genes with modest relative risks. Here, the use of family member controls helped confirm a population association that could otherwise have been dismissed as an artifact of population stratification.

Finally, an example of population stratification in cancer research has recently been presented.<sup>6</sup> Here, the relation between a *CYP3A4* variant and prostate cancer among African-Americans was investigated with a case-control study. Ignoring potential population stratification within this population, a strong positive association between the *CYP3A4* variant and disease was observed ( $P = 0.0007$ ). To correct for potential population stratification, the corresponding *CYP3A4*-prostate cancer test statistic was divided by the average of the test statistics obtained when looking at the relation between 10 unlinked genetic markers and prostate cancer. (This genomic adjustment approach to addressing population stratification is discussed further below.) This correction resulted in the *CYP3A4*-prostate cancer  $P$  increasing to 0.254, suggesting that population stratification may have led to the initial positive result. Of course, there remain some questions surrounding this genomic control approach, such as the type and number of unlinked markers.

Beyond these somewhat extreme examples, a more complete understanding of population stratification requires consideration of whether the circumstances leading to this phenomenon commonly exist. As noted above, the potential for this bias hinges on the heterogeneity in allele frequencies and disease risks across and within populations.

**Population Heterogeneity in Allele Frequencies.** Allele frequencies of many genes have been shown to vary substantially across populations (17). Moreover, the extent of variation is directly related to the genetic distance between populations (18). Because infectious disease exerts strong negative population pressure, often killing people before they reach reproductive age, it is a key factor underlying genetic diversity. Populations have historically been subjected to different infections, depending on the ecology of where they live. Genes that control immune response to infections (different genes, depending on the disease) are therefore the most highly polymorphic and the most subject to confounding by population stratification. Furthermore, because many of the genes regulating immune surveillance may also be involved in other diseases such as cancer, lessons from infectious disease may have broader relevance, even if infectious agents themselves are not involved in the etiology of most cancers (19, 20).

For the purposes of candidate gene association studies, whites of European or Middle Eastern origin are generally regarded as having a relatively homogeneous gene pool, but in fact, there is tremendous variation in genes across Europe, the

Middle East, and India/Pakistan (21). The degree of variation is, of course, dependent on the allele being examined. Alleles that cause rare diseases, such as the cystic fibrosis gene or the *Rh* gene, have relatively homogeneous patterns across Europe (21). On the other hand, genes that are highly polymorphic, such as the HLA alleles, have much greater variation. There are at least four major HLA combination patterns detected throughout Europe, with a concentric gradient spreading out from the Middle East, and various other minor patterns (21). Sokal *et al.* (22) showed that across 3384 localities in Europe, 49 of 59 alleles had substantial variations in allele frequencies. Therefore, populations fairly close together have been shown to exhibit substantial allele frequency variations, leading to potential differences within apparently "homogeneous" groups.

In some parts of the United States, this subracial ethnic variation is not a problem. For example, 80% of the white population in Illinois reported to the 1990 United States Census Bureau that their ancestry was German. Thus, the chance of producing a spurious association in an ethnicity-matched population there is unlikely. Elsewhere, however, such as in California, the origins of individuals within the broadly classified ethnic groups are extremely diverse, so that there is not even a 50% majority of any particular group, and there are large proportions of many disparate groups represented and many mixed-race individuals.

**Population Heterogeneity in Disease Rates.** Of course, for population stratification to occur, the variation in allele frequencies must also be correlated with the subpopulation's variation in their baseline disease rates (*i.e.*, the rates not attributable to the specific candidate gene under study; Ref. 23). In particular, if the baseline rates do not vary by subpopulation, then confounding cannot occur. Nevertheless, there are numerous examples of strong gradients in disease rates across and within countries or major ethnic groups.

Genetic variation is particularly pronounced in infectious disease susceptibility because it is the most genetically dynamic physiological system, having evolved for quick adaptation. Epidemiological studies of infectious diseases indicate that there exists a broad range of susceptibilities to the major infectious diseases, depending on whether populations have been in equilibrium with a disease for a long time or have been exposed only relatively recently. This heterogeneity is apparent even within broad categories of ethnic groups.

Population heterogeneity in disease rates has been particularly well described for cancer and for some sites span a range of 100-fold or more (24). For example, the rate of stomach cancer is much higher among people in Asian countries (*e.g.*, 71–96/100,000 in Japan in comparison with non-Hispanic whites in the United States (5–9/100,000)). Furthermore, disease rates can vary within the broadly defined populations as well. For example, breast cancer incidence rates across Europe range from 26/100,000 in Kielce, Poland to 95/100,000 in Isere, France. Although migrant studies have shown that rates of many cancers tend to converge toward those of their host country in a few generations, suggesting environmental acculturation, their convergence of gene frequencies would be expected to be much slower, depending upon the rate of intermarriage. Thus, although the potential for confounding by unmeasured environmental risk factors is reduced, the potential for confounding by other unmeasured genes remains.

**Confounding.** If there were only two subpopulations and if they differed in both allele frequency and baseline risk, then ignoring these differences in a candidate gene case-control study would lead to confounding. Whether the direction of

<sup>5</sup> The abbreviation used is: TDT, transmission-disequilibrium test.

<sup>6</sup> Rick A. Kittles, Human Genetics, in press, <http://link.springer.de/link/service/journals/00439/contents/tfirst.htm>.

ensuing bias is positive or negative would depend upon whether these differences are in the same or opposite directions. But whatever the direction of the bias, it will always increase the chances of a false-positive significance test if the candidate gene has no causal effect on the risk of disease. The potential magnitude of such confounding bias and distortion of significance levels is well described in standard epidemiological textbooks (25–28). If, however, there are a large number of subpopulations, then even if they differ markedly in both allele frequencies and baseline rates, it seems unlikely that there would be a systematic correlation between these two factors simply by chance (23) that is without some causal reason for such a correlation. In that case, one could argue that that correlation is itself part of the causal pathway that needs to be understood. This argument is based on a combination of simulation studies and empiric analyses, which shed valuable light on the magnitude of the potential problem in practice (23). We have no fundamental quarrel with these conclusions, at least with regard to non-Hispanic whites of European descent. However, we note that not all association studies have been conducted in such populations, and that mixed-race individuals are increasingly common in many parts of the United States. It may be difficult to determine individuals' ethnic admixture in such populations (although the genomic control approaches discussed below may prove helpful). Indeed, some of the "classic" examples discussed above have involved other racial groups and mixtures of two groups, such as the Knowler study of Pima Indians and the Blum study of African Americans and whites.

Even in the absence of confounding bias, population stratification can distort significance levels through "cryptic relatedness" (29–32), *i.e.*, unobserved ancestral relationships between individual cases and controls who are naively treated as independent in the standard  $\chi^2$  test. In particular, pairs of cases are likely to be more closely related than are pairs of controls or case-control pairs if in fact their disease does have a common genetic basis. This will have the effect of inflating the "effective" sample size, thereby increasing the false-positive rate, even in the absence of any confounding bias. However, this effect is likely to be more important in inbred population isolates than in large out-bred populations. Nevertheless, even if the magnitude of the bias attributable to either confounding or cryptic relatedness is small, the effect on significance levels is related to sample size, and hence the very large case-control studies currently being contemplated involving thousands of subjects could have considerably inflated false-positive rates. Although there are thus far no empirical data on the magnitude of the overdispersion of the  $\chi^2$  test of association caused by cryptic relatedness, theoretical arguments, combined with reasonable estimates of the size of the relevant "Wright coefficient of inbreeding" for European populations, led Devlin *et al.* (30, 31) to conclude that "While bias can be a critical factor in traditional epidemiological studies, we argue that overdispersion is the dominant consequence of confounding in genetic studies."

Population stratification also has the potential to confound inferences about gene-environment or gene-gene interactions, although generally to a much lesser extent (23). The reason for this is that population stratification will generally bias the estimate of the effect of a gene to about the same extent in both exposed and unexposed groups, unless there is substantial variation between subpopulations in the association between genotype and environment (although it also depends in part on such other factors as the variability in exposure prevalence, the average magnitude of the genotype-exposure association, and

the association between baseline rates and exposure prevalence; details available from the authors on request).

Of course, some confounding can be controlled by stratification, matching, or statistical adjustment, as discussed further below, but this is generally only possible for major racial groups. What is at issue is how much variation in allele frequencies and disease rates there is between subdivisions of the major ethnic categories. Unfortunately, there are very little data presently available to address this question. This should be a major priority for the next generation of efforts to characterize human genetic variation, as in the planned Human Genome Diversity Project<sup>7</sup> and Environmental Genome Project,<sup>8</sup> for example. However, controversy remains over whether these projects should retain ethnic identifying information.

**Failures to Replicate.** Part of our concern about the potential magnitude of the population stratification problem derives from the widespread, but largely anecdotal, impression that the literature on candidate gene associations is fraught with a disturbingly high frequency of failures to replicate (33–36). For example, Cardon and Bell (35) state that "there are numerous examples of associations that cannot be replicated, which has led to skepticism about the utility of the approach for common conditions." As noted by London *et al.* (37), much of the variability in the reported outcomes, beyond what might be expected just by chance, can be traced to methodologic differences between studies, including such factors as use of inappropriate controls (convenience samples such as lab personnel or hospital controls with other diseases), failure to control even crudely for ethnicity, and multiple significance testing (often with selective reporting of results). The latter can be particularly severe when associations have been identified in a genome-wide scan (38). However, we believe that at least some of the heterogeneity in results could be attributable to residual population stratification. Without more detailed data on the ethnic composition of the subjects in conflicting studies, it is impossible to reach any firm conclusions about whether this is indeed the explanation. Thus, we do not contend that population stratification is the most likely culprit of the failure to replicate, although some authors may have made this interpretation. In an editorial bemoaning that "the majority of association studies are never replicated," *Nature Genetics* issued the following guidelines (39):

*Nature Genetics* continues to welcome submissions of association studies of high quality. Ideally, they should have large sample sizes, small *P*s, report associations that make biological sense and alleles that affect the gene product in a physiologically meaningful way. In addition, they should contain an initial study as well as an independent replication, the association should be observed both in family based and population-based studies, and the odds ratio and/or attributable risk should be high. Few studies will meet all criteria, but to minimize our risk, we will apply high standards. In general, we will expect manuscripts reporting genetic associations to include an estimate of the effect size and to contain either a replication in an independent sample or physiologically meaningful data supporting a functional role of the polymorphism in question. (emphasis added)

How widespread is this problem of nonreplication? In an attempt to address this question, Terwilliger and Weiss (40) reviewed 18 months of reports in the journals *Neuropsychiatric Genetics* and *Psychiatric Genetics* and plotted the distribution

<sup>7</sup> Internet address: <http://www.stanford.edu/group/morrinst/hgdp.html>.

<sup>8</sup> Internet address: <http://www.niehs.nih.gov/envgenom/home.htm>.

of  $P$ s for reported candidate gene associations. Under the null hypothesis that there are no true positive associations in the entire ensemble [assuming that all associations tested were published, *i.e.*, no “publication bias” (41)], the distribution of  $P$ s should be uniform between 0 and 1 and departures from uniformity can be used to derive an estimate of the number of true-positive associations (42). In this analysis, Terwilliger and Weiss were unable to reject the global null hypothesis of uniformity, suggesting that, at least within these journals, there may indeed be a problem with replicating results. Of course, it is a theoretical possibility that the global null hypothesis is true, but given the diversity of genes and outcomes under study, many with strong prior basis, it seems unlikely to us that all of the null hypotheses could be true.

We are aware of only two other such systematic surveys of replication rates for candidate gene associations. Ionnidis *et al.* (43) conducted a meta-analysis of 370 studies concerning 36 associations and found significant heterogeneity for the majority of the reported associations; in particular, they concluded that “the results of the first study correlate only modestly with subsequent research on the same association.” Hirschhorn *et al.* (44) identified 162 associations between diseases and common polymorphisms that have been studied three or more times. Of these, only 6 were highly consistently reproducible (75% or more of studies positive); of the remaining 156, 98 had at least one more positive study (the “mixed” category), and 58 had only one positive study.

What might one expect in terms of the frequency of true positive associations, purely on the basis of significance levels? True associations are probably rare, given the 30,000–40,000 genes and >1,000,000 polymorphisms in the genome. Even if one guessed well as to “candidate” genes, one has to admit that any given polymorphism has a very small chance of being a causal allele. Thus, even with high specificity, false positives will far outweigh true positives. For example, assume one adopts a conventional significance level of  $\alpha = 5\%$  and only conducts studies with at least 50% power. If there were 1,000 candidate alleles to choose from, of which say 10 were really causal, then one would expect 50 false positives ( $1,000 \times 5\%$ ), which would likely be many more than the expected number of true positives ( $5 = 10 \times 50\%$ ). On this basis, it seems likely that false positives will be a much bigger problem than false negatives, suggesting that an extremely conservative significance level should be adopted when testing many candidate gene associations, unless one has a strong prior basis for belief in any particular one. The problem of multiple comparisons has been widely discussed in both the epidemiological and genetic literatures (45–48) but is somewhat beyond the scope of this article. In any event, the rate of false positives that would be generated by the action of chance alone would be expected to be the similar for case-control designs using unrelated or family-based controls.

The reader might of course wonder whether the situation is any better for gene mapping studies by linkage analysis, where there are also numerous examples of replication failure, despite the well-developed theory for inference based on sequential testing underlying the conventional criterion of a lod score of 3 (49) or the more recently suggested criterion of 3.6 (46). Unfortunately, we were unable to identify any recent systematic studies of replication rates for linkage findings, although another recent *Nature Genetics* editorial (50) echoed the prevailing sentiment that “Genome-wide linkage scans designed to localize disease genes have yielded few significant findings, and failure to reproduce published linkage results is endemic” (emphasis added). However, a review of 1665 marker-marker

linkages conducted before modern DNA markers were available, Rao *et al.* (51) found that lod scores of >1 were seldom replicated, whereas those >2 usually were.

### Approaches to Dealing with Confounding

**Better Measures of Populations.** The conventional approaches to dealing with confounders in epidemiology are by matching, stratification, or multivariate adjustment models. In the context of population stratification, this calls for more detailed information on ethnicity than such broad conventional categories as Caucasian, African American, Hispanic, Asian, and Other. There are two dimensions to this challenge: (a) individuals must be allocated to the finest ethnic origin categories that can reliably be determined; and (b) individuals from mixed-ethnicity families must be treated appropriately.

Definition of suitable ethnicity categories is a matter of judgment, entailing consideration of the trade-off between specificity and reliability. In general, however, our view is that the investigator is better off collecting the information with a high degree of specificity, reasoning that even if there is some misclassification, it is probably less than that resulting from use of overly broad categories to begin with. If cases and controls are individually matched, as is generally desirable to allow for various confounding variables in addition to ethnicity, it may not always be possible to obtain an exact match on ethnicity, but at least an approximate match should be attempted, with further adjustments made in the analysis. Because measurement error in a stratifying variable will generally lead to partial loss of control of confounding (52), one might consider an analysis that allows for the uncertainties in the measurement of ethnicity, in the spirit of some approaches to correction for exposure measurement error (53), although these uncertainties can themselves be difficult to quantify.

Mixed-ethnicity families pose greater challenges. Here, it can be very difficult, if not impossible, to find a matching control for an individual with multiple ethnic origins, and the investigator is forced to rely on multivariate models for adjustment. The key is to obtain as detailed information as possible on the ethnic origins of the subjects' ancestors. We generally recommend that questions be asked not simply about the subjects' own ethnic identification but also about the origins of his/her parents' and, if possible, grandparents [see, for example, Whittemore *et al.* (54), who inquired about the ethnic origins of parents and grandparents]. In fact, it is probably worthwhile to query cases and controls about all of their known ancestor's countries of origin. Rather than allocate the entire individual to a single stratum in the analysis, as is conventionally done, one can construct a covariate for each stratum, giving the proportion of ancestors derived from each ethnic group and then include these covariates as adjustment variables in a multiple logistic regression model.

**Use of Family-Member Controls.** The difficulty of finding ethnically matched controls can be completely overcome by use of family-member controls. The most commonly used familial case-control designs involve the use of siblings or parents as controls (55–57). Sibling controls are derived from exactly the same gene pool as the cases and thus represent exactly matched controls, but they pose other practical and statistical difficulties. The major practical difficulty is that not every case will have an available sibling; if sibship size or other determinants of availability are associated with genotype, selection bias will result, which could go in either direction, depending upon the direction of the association, and could increase the risk of spurious associations with candidate genes that are associated with such

selective factors. A second difficulty is that controls should generally be selected from siblings who have already survived to the age at diagnosis of the case free of the disease (58). In practice, this will generally tend to limit control eligibility to older siblings, which can lead to confounding by factors related to year of birth, family size, or birth order, particularly if time-dependent exposure factors are also under consideration. Siblings are also more likely to have the same genotype as the case than are unrelated controls, thereby leading to some loss of statistical efficiency (*i.e.*, larger sample sizes required to attain the same statistical precision). They are also more likely to have common environments, leading to some loss of efficiency for the main effect of environment as well, although surprisingly the use of sibling controls generally *improves* efficiency for gene-environment interactions (57, 59).

As an alternative to siblings, one might consider using cousins as controls, which are less likely to be genotype concordant and more readily available, allowing for closer matching on age and year of birth. However, in contrast with siblings, cousins may be more difficult to identify, less motivated to participate, and do not provide the same protection from ethnic stratification because they have only one set of grandparents in common. They may also introduce geographic confounding if cases are limited to a defined region and cousin controls live elsewhere (60). An advantage to family-based designs, however, can be the power gain by restricting to multiple-case families, particularly for rare alleles (56), although care is needed in the analysis to provide unbiased tests and estimates.

An increasingly common study design uses parents to form controls. More precisely, it is not the parents themselves who are the controls in this design but the set of genotypes the parents could have transmitted to the case, given their own genotypes (the case's "pseudosibs"). This design can be analyzed as a 1:3 matched case-control study by conditional logistic regression, as described by Self *et al.* (61). (From the perspective of statistical efficiency, the effective matching ratio is 1: $\infty$ .) A special case of this analysis is the TDT (16), in which the unit of analysis is not the subjects' genotypes but their two alleles, one from each parent, each being compared with the nontransmitted allele. This test is formally equivalent to the score test from the Self *et al.* (61) likelihood under a multiplicative model for penetrance, whereas the Self *et al.* (61) approach allows one to test hypotheses about alternative modes of inheritance. The use of pseudosib controls has better statistical efficiency than sibling or cousin controls (even more than population controls for a recessive gene), but the requirement that parents be available for genotyping limits its usefulness for late-onset diseases, such as most cancers. Pseudosib controls are generally slightly less efficient than population controls for estimating  $G \times E$  interactions, except under a recessive model. Witte *et al.* (57) further explore the theoretical, statistical, and practical considerations in choosing between unrelated population or family controls.

**Genomic Adjustment.** Recently, a number of authors have proposed using genomic information to help address the problem of bias attributable to population stratification and overdispersion attributable to cryptic relatedness. In particular, with a panel of polymorphic markers that are not linked to the candidate gene under study, one can attempt to address the issue population stratification by: (a) using an overdispersion model to determine a test statistic's appropriate empirical distribution; (b) evaluating whether stratification exists; and (c) using a latent-class model to distinguish homogeneous subpopulations. The restriction to unlinked genes is intended to

avoid the loss of statistical precision attributable to over adjustment by a correlate of the genes of interest; it is not necessary or desirable, however, to exclude all markers that are associated with the candidate gene but not linked with them, because it is precisely these associations that are most informative to control for the cryptic stratification.

With regard to the overdispersion approach for addressing population stratification, Devlin and Roeder (29), Bacanu *et al.* (62), and Reich and Goldstein (63) point out that in the presence of population stratification, the null distribution of the usual  $\chi^2$  test of the hypothesis of no association between the disease and a candidate gene will tend to be shifted toward higher values, leading to an inflated false-positive rate. If one had enough unlinked markers that were not causally related to the disease, one could in principle estimate this null distribution simply by tabulating the distribution of the observed test statistics across all of the markers and comparing the value for the specific candidate gene to this empiric distribution. In practice, however, the number of "null" markers that would be required to determine significance levels with any precision would be prohibitive, probably even with modern single nucleotide polymorphism array technologies. However, the theoretical null distribution in the presence of stratification is shifted by a multiplicative constant that can be estimated in a straightforward manner from the observed  $\chi^2$  values for the null marker data in the same subjects in which one wishes to test a specific candidate gene (29, 62, 63). For reliable estimation, one may require 50 or more null markers (30). This genomic adjustment approach is computationally simple and simultaneously addresses issues of population stratification bias and overdispersion attributable to cryptic relatedness; furthermore, it allows for a large number of potential subgroups (*i.e.*, it works well with very fine-scale substructure; Ref. 29). An investigation of power indicates that the genomic control approach can generally be more powerful than the TDT (63). It can be undertaken with pooled DNA samples as well, which can be substantially less expensive than individual genotyping.

In another approach, Pritchard and Rosenberg (64) suggest that inference follow a two-step process; first one uses a panel of markers to test for stratification and then proceeds to evaluating the candidate gene association only if homogeneity is not rejected. By simulation, they have explored the critical values that should be used at each stage of the inference process and shown that the method performs well using a panel of a couple dozen markers. However, they provide no guidance for what the investigator should conclude if the hypothesis of homogeneity is rejected in the first stage. Most epidemiologists would be unhappy to learn that a study they had laboriously conducted should simply be discarded because of the existence of population stratification at a panel of markers in which they had no particular interest!

The third, latent-class approach, entails using genomic information to distinguish subpopulations within which any potential population stratification is minimized. For example, following the above work, Pritchard *et al.* (65) developed a Bayesian approach to estimation of ethnic origins. In a population comprising a mixture of an unknown number of subpopulations without admixture, their approach leads to an estimate of the posterior distribution of the number of such subpopulations and the probability that each individual belongs to any given subpopulation. In the presence of admixture, the method estimates the proportion of an individual's genome that derives from each subpopulation. Schork *et al.* (66) have recently proposed a similar approach for addressing population stratification. Kim *et al.* (67) presented related work suggesting

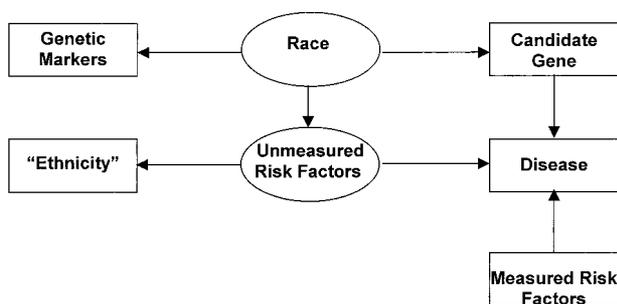


Fig. 1. Schematic representation of conceptual model for population stratification. Boxes represent observed variables; ellipses, unobserved variables. We assume that the candidate gene and the unobserved risk factors are conditionally independent, given race. Genetic markers and self-reported ethnicity are viewed as surrogates for race and risk factors, respectively.

using phylogenetic trees defined by nonlinked single nucleotide polymorphisms to cluster individuals into homogeneous subgroups. For all of these approaches, once one has an estimate of the probability of membership within each subgroup, this information can be incorporated into the analysis of a case-control study to obtain an estimate of the candidate gene's effect, adjusted for potential stratification (68). Bacanu *et al.* (69) have also generalized the approach to quantitative traits. Satten *et al.* (70) have recently described a different approach to adjusting for population stratification using molecular markers, based on a maximum likelihood approach to latent class models, summing each individual's likelihood contribution overall possible strata. The naïve alternative of adding all of the markers as covariates in a multiple logistic regression, although appealing for its simplicity, has been shown to produce less precise estimates than the Pritchard *et al.* (65) approach.<sup>9</sup> Pritchard and Donnelly (32) provide simulation studies comparing the Bayesian clustering and latent-class approaches.

Of course, other aspects of ethnicity (cultural and environmental) may play a more important role in the variation in cancer rates than genetics (23). Nevertheless, to control for confounding it is sufficient to control for the determinants of gene frequency. Our conceptual framework is shown in Fig. 1. We postulate an unobserved stratifying factor we shall call "race," which is a determinant of allele frequencies at the candidate gene (and at each of the marker loci), and is also a determinant of one or more unobserved risk factors (other causal genes, environmental, behavioral, or cultural influences). If, conditional on race, the candidate gene and the other risk factors are independent, then an analysis that adjusts for race but not the other risk factors would give unbiased estimates of the relative risk for the candidate gene. Likewise, an analysis that adjusts for the other risk factors but not race would also be unbiased. However, because neither race nor the other risk factors are directly observable, it is necessary to adjust for surrogates for one or the other. The question thus reduces to whether "genomic control" is a better surrogate for race than self-reported ethnicity is for the unobserved risk factors.

### Summary and Conclusions

In light of the classic examples of population stratification, as well as the population heterogeneity in allele frequencies and

disease rates, we believe that population stratification is a sufficiently serious concern to merit careful consideration in interpreting the results of any candidate-gene association that is not based on the use of family-member controls. This concern would be mitigated by careful attention to the standard principles of epidemiological study design, including the choice of a mechanism for selecting controls that are representative of the source population of cases and by making some effort to control ethnicity by restriction, matching, or stratified or multivariate analysis. For any of these approaches to be successful, information on ethnic origin should be obtained in the greatest detail that is practically feasible. Although there will always be some risk of residual confounding by ethnicity in case-control studies using unrelated controls, we do not feel that this problem is qualitatively different from the problem of uncontrolled confounding in virtually any other observational epidemiological study. However, the considerable range of variation in allele frequencies across and within ethnic groups and the enormous magnitude of some genetic risks (compared with those from most environmental agents) suggests that the problem of confounding by population stratification could be more serious in magnitude than in traditional environmental epidemiology. The availability of family-based case-control designs that completely eliminate these concerns suggests that no candidate gene association should be considered "confirmed" until replicated by such a study, or at least by multiple well-designed studies in different populations where any effects of population stratification or other methodologic biases are unlikely to act in a consistent manner.

Rather than continued debate about whether population stratification is or is not a serious concern, we call for a systematic program of research to understand the magnitude of the problem in general. Additional studies of the variation in allele frequencies of many genes and of the variation in baseline rates for a wide range of disease and the correlation between the two would be extremely useful, as would further meta-analyses of candidate gene associations, with particular attention to study of the determinants of observed heterogeneity in findings (71). Finally, we suggest that the promising approaches to use of genomic information as a means of detecting and controlling for population stratification merit further application and development.

### Acknowledgments

We thank Dr. Wendy Cozen for helpful discussions.

### References

- Lander, E. S., and Schork, N. J. Genetic dissection of complex traits. *Science* (Wash. DC), 265: 2037–2048, 1994.
- Ewens, W. J., and Spielman, R. S. The transmission/disequilibrium test: history, subdivision, and admixture. *Am. J. Hum. Genet.*, 57: 455–464, 1995.
- Altschuler, D., Kruglyak, L., and Lander, E. S. Genetic polymorphism and disease. *N. Engl. J. Med.*, 338: 1626, 1998.
- Witte, J. S., Gauderman, W. J., and Thomas, D. C. Population stratification in association studies. *Genet. Epidemiol.*, 15: 538, 1998.
- Khoury, M. J., and Beaty, T. H. Applications of the case-control method in genetic epidemiology. *Epidemiol. Rev.*, 16: 134–150, 1994.
- Khoury, M. J., and Yang, Q. The future of genetic studies of complex human diseases: an epidemiologic perspective. *Epidemiology*, 9: 350–354, 1998.
- Wacholder, S., Rothman, N., and Caporaso, N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiologic studies of common polymorphisms and cancer. *Cancer Epidemiol. Biomark. Prev.*, 11: this issue, 2002.
- Zheng, C., and Elston, R. C. Multipoint linkage disequilibrium mapping with particular reference to the African-American population. *Genet. Epidemiol.*, 17: 79–101, 1999.

<sup>9</sup> J. S. Witte, unpublished data.

9. Knowler, W. C., Williams, R. C., Pettitt, D. J., and Steinberg, A. G. *Gm<sup>3,5,13,14</sup>* and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am. J. Hum. Genet.*, *43*: 520–526, 1988.
10. Gelernter, J., Goldman, D., and Risch, N. The *A1* allele at the D2 dopamine receptor gene and alcoholism. A reappraisal. *J. Am. Med. Assoc.*, *269*: 1673–1677, 1993.
11. Blum, K., Noble, E. P., Sheridan, P. J., Montgomery, A., Ritchie, T., Jagadeeswaran, P., Nogami, H., Briggs, A. H., and Cohn, J. B. Allelic association of human dopamine D2 receptor gene in alcoholism. *J. Am. Med. Assoc.*, *263*: 2055–2060, 1990.
12. Edenberg, H. J., Foroud, T., Koller, D. L., Goate, A., Rice, J., Van Eerdewegh, P., Reich, T., Cloninger, C. R., Nurnberger, J. I. J., Kowalczyk, M., Wu, B., Li, T.-K., Conneally, P. M., Tischfield, J. A., Wu, W., Shears, S., Crowe, R., Hesselbrock, V., Schuckit, M., Porjesz, B., and Begleiter, H. A family-based analysis of the association of the dopamine D2 receptor (DRD2) with alcoholism. *Alcohol Clin. Exp. Res.*, *22*: 505–512, 1998.
13. Blomqvist, O., Gelernter, J., and Kranzler, H. R. Family-based study of *DRD2* alleles in alcohol and drug dependence. *Am. J. Med. Genet.*, *96*: 659–664, 2000.
14. Bell, G., Horita, S., and Karam, J. H. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes*, *33*: 176–183, 1984.
15. Spielman, R. S., Baur, M. P., and Clerget-Darpoux, F. Genetic analysis of IDDM: summary of GAW5-IDDM results. *Genet. Epidemiol.*, *6*: 43–58, 1989.
16. Spielman, R. S., McGinnis, R. E., and Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, *52*: 506–516, 1993.
17. Perez-Lezaun, A., Calafell, F., Mateu, E., Comas, D., Bosch, E., and Bertranpetit, J. Allele frequencies for 20 microsatellites in a worldwide population survey. *Hum. Hered.*, *47*: 189–196, 1997.
18. Goddard, K., Hopkins, P. J., Hall, J. M., and Witte, J. S. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.*, *66*: 216–234, 2000.
19. Boon, T., Cerottini, J. C., Van den Eynde, B., Van der Bruggen, P., and Van Pel, A. Tumor antigens recognized by T lymphocytes. *Annu. Rev. Immunol.*, *12*: 337–366, 1994.
20. McMichael, A. J. In: A. J. McMichael and W. F. Bodmer (eds.), *A New Look at Tumor Immunology*, pp. 5–21. Plainview, NY: Cold Spring Harbor Laboratory, 1992.
21. Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. *The history and geography of human genes*. Princeton, NJ: Princeton University Press, 1994.
22. Sokal, R. R., Harding, R. M., and Oden, N. L. Spatial patterns of human gene frequencies in Europe. *Am. J. Phys. Anthropol.*, *80*: 267–294, 1989.
23. Wacholder, S., Rothman, N., and Caporaso, N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J. Cell. Biochem. Suppl.*, *92*: 1151–1158, 2000.
24. Ferlay, J., Black, R. J., Whelan, S. L., and Parkin, D. M. C15VII: Electronic Database of Cancer Incidence in Five Continents, Vol. VII. Lyon, France: IARC Scientific Publications, 1997.
25. Breslow, N. E., and Day, N. E. *Statistical methods in cancer research. I. The Analysis of Case-Control Studies*. Statistical Methods in Cancer Research. Lyon, France: IARC Scientific Publications, 1980.
26. Klienbaum, D. G., Kupper, L. L., and Morgenstern, H. *Epidemiologic Research: Principles and Quantitative Methods*. Belmont, CA: Lifetime Learning Publications, 1982.
27. Miettinen, O. S. *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*. New York: John Wiley & Sons, 1985.
28. Rothman, K. J., and Greenland, S. *Modern epidemiology*. Philadelphia: Lippincott-Raven, 1998.
29. Devlin, B., and Roeder, K. Genomic control for association studies. *Biometrics*, *55*: 997–1004, 1999.
30. Devlin, B., Roeder, K., and Bacanu, S.-A. Unbiased methods for population-based association studies. *Genet. Epidemiol.*, *21*: 273–284, 2001.
31. Devlin, B., Roeder, K., and Wasserman, L. Genomic control, a new approach to genetic-based association studies. *Theor. Pop. Biol.*, *60*: 155–166, 2001.
32. Pritchard, J. K., and Donnelly, P. Case-control studies of association in structured or admixed populations. *Theor. Pop. Biol.*, *60*: 227–237, 2001.
33. Vineis, P., Malats, N., Lang, M., d'Errico, A., Caporaso, N., Cuzick, J., and Boffetta, P. Metabolic polymorphisms and susceptibility to cancer. Lyon, France: IARC Scientific Publications, 1999.
34. Altschuler, D., Hirschhorn, J. N., Klannemark, M., Lindgren, C. M., Vohl, M., Nemesh, J., Lane, C. R., Schaffner, S. F., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., Hudson, T. J., Daly, M., Groop, L., and Lander, E. S. The common PPAR $\gamma$  Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.*, *26*: 76–80, 2000.
35. Cardon, L. R., and Bell, J. I. Association study designs for complex diseases. *Nat. Rev. Genet.*, *2*: 91–99, 2001.
36. Dahlman, I., Eaves, I. A., Kosoy, R., Morrison, V. A., Heward, J., Gough, S. C. L., Allahabadi, A., Franklyn, J. A., Tuomilehto, J., Tuomilehto-Wolf, E., Cucca, F., Guja, C., Ionescu-Tirgoviste, C., Stevens, H., Carr, P., Nutland, S., McKinney, P., Shield, J. P., Wang, W., Cordell, H. J., Walker, N., Todd, J. A., and Concannon, P. Parameters for reliable results in genetic association studies in common disease. *Nat. Genet.*, *30*: 149–150, 2002.
37. London, S., Daly, A., Thomas, D., Caporaso, N., and Idle, J. Methodological issues in the interpretation of studies of the CYP2D6 genotype in relation to lung cancer risk. *Pharmacogenetics*, *4*: 107–108, 1994.
38. Goring, H. H., Terwilliger, J. D., and Blangero, J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Genet.*, *69*: 1357–1369, 2001.
39. Freely associating. *Nat. Genet.*, *22*: 1–2, 1999.
40. Terwilliger, J. D., and Weiss, K. M. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.*, *9*: 578–594, 1998.
41. Begg, C. B., and Berlin, J. A. Publication bias: a problem in interpreting medical data. *J. R. Statist. Soc. Ser. A*, *151*: 419–463, 1988.
42. Schweder, T., and Spotjvol, E. Plots of *P*-values to evaluate many tests simultaneously. *Biometrika*, *69*: 493–502, 1982.
43. Ionnisidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ionnisidis, D. G. Replication validity of genetic association studies. *Nat. Genet.*, *29*: 306–309, 2001.
44. Hirschhorn, J. N., Lohmueller, K., Byrne, E., and Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.*, *4*: 45–61, 2002.
45. Greenland, S. Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary-testing, and empirical-Bayes regression. *Stat. Med.*, *12*: 717–736, 1993.
46. Lander, E. S., and Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.*, *11*: 241–247, 1995.
47. Thomas, D. C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M., and Armstrong, B. G. The problem of multiple inference in studies designed to generate hypotheses. *Am. J. Epidemiol.*, *122*: 1080–1095, 1985.
48. Witte, J., Elston, R., and Schork, N. Genetic dissection of complex traits. *Nat. Genet.*, *12*: 355–356, 1996.
49. Morton, N. E. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.*, *7*: 277–318, 1955.
50. Altshuler, D., Daly, M., and Kruglyak, L. Guilt by association. *Nat. Genet.*, *26*: 135–137, 2000.
51. Rao, D. C., Keats, B. J., Morton, N. E., Yee, S., and Lew, R. Variability of human linkage data. *Am. J. Hum. Genet.*, *30*: 516–529, 1978.
52. Greenland, S. The effect of misclassification in the presence of covariates. *Am. J. Epidemiol.*, *112*: 564–569, 1980.
53. Thomas, D. C., Stram, D., and Dwyer, J. Exposure measurement error: influence on exposure-disease relationships and methods of correction. *Annu. Rev. Publ. Health*, *14*: 69–93, 1993.
54. Whittemore, A., Kolonel, L., and Wu, A. Prostate cancer in relation to diet, physical activity, and body size in blacks, whites, and Asians in the United States and Canada. *J. Natl. Cancer Inst.*, *87*: 652–661, 1995.
55. Schaid, D. J., and Rowland, C. Use of parents, sibs and unrelated controls for detection of associations between genetic markers and disease. *Am. J. Hum. Genet.*, *63*: 1492–1506, 1998.
56. Teng, J., and Risch, N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res.*, *9*: 234–241, 1999.
57. Witte, J. S., Gauderman, W. J., and Thomas, D. C. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am. J. Epidemiol.*, *148*: 693–705, 1999.
58. Lubin, J. H., and Gail, M. H. Biased selection of controls for case-control analysis of cohort studies. *Biometrics*, *40*: 63–75, 1984.
59. Gauderman, W., Witte, J., and Thomas, D. Family-based association studies. *Monogr. Natl. Cancer Inst.*, *26*: 31–37, 1999.
60. Sigmund, K. P., and Langholz, B. Ascertainment bias in family-based case-control studies. *Am. J. Epidemiol.*, *155*: 857–880, 2002.
61. Self, S. G., Longton, G., Kopecky, K. J., and Liang, K. Y. On estimating HLA/disease association with application to a study of aplastic anemia. *Biometrics*, *47*: 53–61, 1991.
62. Bacanu, S. A., Devlin, B., and Roeder, K. The power of genomic control. *Am. J. Hum. Genet.*, *66*: 1933–1944, 2000.

63. Reich, D. E., and Goldstein, D. B. Detecting association in a case-control study while correcting for population stratification. *Genet. Epidemiol.*, 20: 4–16, 2001.
64. Pritchard, J. K., and Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.*, 65: 220–228, 1999.
65. Pritchard, J. K., Stephens, M., and Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics*, 155: 945–959, 2000.
66. Schork, N. J., Fallin, D., Thiel, B., Xu, X., Broeckel, U., Jakob, H. J., and Cohen, D. The future of genetic case-control studies. *In*: D. C. Rao and M. Province (eds.), *Genetic Dissection of Complex Traits*, pp. 191–212. San Diego: Academic Press, 2001.
67. Kim, L.-L., Fijal, B. A., and Witte, J. S. Hierarchical modeling of the relation between sequence variants and a quantitative trait: addressing multiple comparison and population stratification issues. *Genet. Epidemiol.*, 21: S668–S673, 2001.
68. Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.*, 67: 170–181, 2000.
69. Bacanu, S.-A., Devlin, B., and Roeder, K. Association studies for quantitative traits in structured populations. *Genet. Epidemiol.*, 22: 78–93, 2002.
70. Satten, G. A., Flanders, W. D., and Yang, Q. 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, 2001.
71. Greenland, S. Quantitative methods in the review of epidemiologic literature. *Epidemiol. Rev.*, 9: 1–30, 1987.