Am. J. Hum. Genet. 65:220–228, 1999

Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies

Jonathan K. Pritchard[1,2] and Noah A. Rosenberg[1]

[1]Department of Biological Sciences, Stanford University, Stanford, and [2]Department of Statistics, University of Oxford, Oxford

Summary

We examine the issue of population stratification in association-mapping studies. In case-control studies of association, population subdivision or recent admixture of populations can lead to spurious associations between a phenotype and unlinked candidate loci. Using a model of sampling from a structured population, we show that if population stratification exists, it can be detected by use of unlinked marker loci. We show that the case-control–study design, using unrelated control individuals, is a valid approach for association mapping, provided that marker loci unlinked to the candidate locus are included in the study, to test for stratification. We suggest guidelines as to the number of unlinked marker loci to use.

Introduction

Association mapping is used to identify chromosomal regions containing disease-susceptibility loci or loci involved in other phenotypic traits of interest. A powerful technique, it has been advocated as the method of choice for mapping complex-trait loci (Risch and Merikangas 1996). The basic approach is to identify marker loci at which some alleles are more frequent among affected individuals (cases) than among unaffected individuals (controls). A statistical association between genotypes at the marker locus and the phenotype is usually thought to imply physical linkage between the marker locus and a disease locus.

This conclusion is reasonable in a randomly mating population, since linkage disequilibrium between unlinked markers breaks down very rapidly over time. However, when population subdivision is present, it is possible to find statistical associations between a disease phenotype and arbitrary markers that have no physical linkage to causative loci (e.g., see Lander and Schork 1994; Ewens and Spielman 1995). Such associations occur because population subdivision (or any other form of nonrandom mating) permits marker-allele frequencies to vary among segments of the population, as the result of genetic drift or founder effects (Slatkin 1991). A disease that is most prevalent in one subpopulation will be associated with *any* alleles that are in high frequency in that subpopulation. Of course, nothing is special about the disease in this regard. Any marker locus that has different allele frequencies in the different subpopulations will be in "linkage" disequilibrium with other markers from throughout the genome.

In this article, we use the term "spurious association" to describe an association between a phenotype and a marker locus when the marker is unlinked to any causative loci. Spurious associations as a result of population subdivision can occur when the sampling is done without regard to ethnicity and the resulting case and control samples contain different frequencies of each ethnic group. We present a model in which disease frequencies differ among subpopulations, causing some subpopulations to be overrepresented in the affected group. In some situations it may be easy to detect this form of stratification, simply by asking each member of the sample to identify his or her ethnicity. However, it is an open question as to whether relatively fine distinctions should be of concern: for example, is stratification in a sample of Irish Americans, Italian Americans, and Jewish Americans likely to produce spurious associations? As discussed below, the severity of the problem of spurious association increases with sample size; thus, perhaps such stratification *will* be an issue in the large studies that will be necessary to identify disease loci with low relative risks.

Spurious associations can also arise in recently admixed populations. For example, association studies of type II diabetes in Pima Indians (who have high rates of diabetes) were flawed because Pima individuals with a high degree of Caucasian ancestry had lower diabetes susceptibility. Hence, any marker loci that were at higher frequency in the Pima than in Caucasians were "asso-

Received December 1, 1998; accepted for publication April 21, 1999; electronically published June 7, 1999.

Address for correspondence and reprints: Dr. Jonathan Pritchard, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1-3TG United Kingdom. E-mail: pritch@stats.ox.ac.uk

0002-9297/99/6501-0028$02.00

ciated" with the disease (Lander and Schork 1994). Ewens and Spielman (1995) describe a model of admixed populations in which spurious associations can arise whenever the frequency of the disease allele differs among the parent populations. Of course, random mating and recombination break down false associations within just a few generations of the initial admixture.

In response to the problem of spurious associations (Falk and Rubinstein 1987; Thomson et al. 1989), Spielman et al. (1993) proposed the transmission disequilibrium test (TDT), which uses the genotypes of parents of affected individuals. The TDT checks for a difference in the transmission probabilities of the candidate alleles from heterozygous parents. Although this approach takes advantage of population-level associations, the TDT is not susceptible to spurious associations that result from stratification.

Although very elegant, the TDT design is usually more labor intensive than a simple case-control design that uses affected individuals and unrelated controls. It may take considerable effort, or may even be impossible, to collect DNA samples from the parents of probands, particularly for late-onset diseases. It may also be difficult to collect DNA from other relatives of probands for which TDT-like statistics have been proposed (Boehnke and Langefeld 1998; Lazzeroni and Lange 1998; Spielman and Ewens 1998).

For this reason the simple case-control approach would often be an attractive study design, were it not for the problem of spurious associations due to population stratification. In the remainder of this article, we argue that the case-control design *can* be a valid test for association, provided that it includes an explicit test for stratification.

Detection of Stratification

It is well known that allele frequencies at random marker loci may differ among ethnic groups (reviewed at length by Cavalli-Sforza et al. 1994). For this reason, if the cases and controls in a gene-mapping study contain different mixtures of ethnic groups, we should expect to find a consistent pattern of allele-frequency differences between cases and controls, at many loci throughout the genome. By contrast, if the cases and controls are well matched, significant allele-frequency differences should be located near disease-susceptibility loci only.

At this time, most association studies examine only a few candidate loci. With use of that approach, if one or more markers show statistical associations, the possibility that the associations are due to population stratification cannot be eliminated.

However, it *would* be possible to detect stratification by typing additional unlinked markers. If stratification is present, then the unlinked markers should also show associations with the phenotype. This idea forms the basis of our test for stratification.

In this article we describe samples as "stratified" if the cases and controls are mismatched; that is, they contain different proportions of each ethnic group, or, in the case of admixture, they contain different fractions of ancestry from each ancestral subpopulation. We would not consider a sample to be stratified if its cases and controls contain *equal* proportions of each of a series of ethnic groups, since such a situation would not be expected to lead to an excessive rate of false positives.

Study Design

We envision a study design comprising two random samples of unrelated individuals: one sample composed of affected individuals and one sample composed of unaffected individuals. Ideally, these samples should be ethnically matched, as well as possible, in advance.

In the next section, we focus on a situation in which markers are chosen at a few candidate loci. The allele frequencies at those markers are tested for association with the phenotype. If one or more of the candidates are significantly associated with the phenotype, we advocate typing an additional set of random markers unlinked to any candidate loci, to test for stratification. If the unlinked markers do not indicate stratification, then the associations can be considered valid. We discuss how many markers should be used.

It seems likely that the recent development of microarray technology for rapid screening of single-nucleotide variation will soon permit a second genotyping strategy: fine-scale genome screens for association, without prespecified candidate loci (Wang et al. 1998). In the Discussion, we also consider tests for stratification in this context.

Tests of Association by Use of Unrelated Controls

Consider a case-control experiment with m_{h} healthy individuals and m_{d} diseased individuals. All individuals are unrelated. We wish to test whether a particular allele, *A*, at the candidate locus, is associated with the disease phenotype (table 1), in the sense that the probability of carrying allele *A* is not independent of case-control status. This can done with the usual Pearson χ^2 test statistic, in the form

$$\chi^2 = (\hat{q}_{\mathrm{d}} - \hat{q}_{\mathrm{h}})^2 8m \frac{m_{\mathrm{d}} m_{\mathrm{h}}}{n_A n_{A^*}}, \qquad (1)$$

where $\hat{q}_{\mathrm{d}}$ and $\hat{q}_{\mathrm{h}}$ are the estimated frequencies of *A* alleles, among affected and healthy individuals, respectively; m is the total number of individuals sampled; and n_A and n_{A^*} are the total numbers of *A* alleles and non-*A* alleles

Table 1

Contingency Table of the Observed Frequencies of *A* Alleles and non-*A* Alleles in Diseased and Healthy Individuals

Alleles	Frequency of [a]		Total No. of Alleles[b]
	A	A*	
Diseased	$\hat{q}_d$	$1-\hat{q}_d$	$2m_d$
Healthy	$\hat{q}_h$	$1-\hat{q}_h$	$2m_h$
Total no. of alleles	n_A	n_{A^*}	$2m$

[a] A^* denotes non-*A* alleles; $\hat{q}_d$ and $\hat{q}_h$ are the observed frequencies of *A* alleles among diseased and healthy individuals, respectively; and n_A and n_{A^*} denote the numbers of *A* and non-*A* alleles.

[b] $2m_d$ and $2m_h$ denote the numbers of alleles in diseased and healthy individuals.

observed (Sokal and Rohlf 1995, eq. [17.13]). The factor of 8 arises because m, m_d, and m_h are numbers of individuals, not numbers of alleles. Under the null hypothesis of no association, $E(\hat{q}_d - \hat{q}_h) = 0$, and X^2 has a χ_1^2 distribution. As described in the example below, this expectation may not be 0 if there is population structure.

We describe a method for determining whether an inferred association may have arisen as the result of sampling from a structured population. To make the discussion concrete, we introduce a model of stratified sampling.

A Model of Stratified Sampling

Suppose that each individual sampled is actually a member of one of n subpopulations but that we select individuals without regard to their origin. Let the probability of sampling an individual from subpopulation i be γ_i (where $\sum_{i=1}^n \gamma_i = 1$), and let the frequency of the disease in the ith subpopulation be p_i. Then, with use of Bayes' rule, the probability (f_i) that an affected individual is from the ith subpopulation is

$$f_i = \frac{\gamma_i p_i}{\sum_{j=1}^n \gamma_j p_j} \quad . \tag{2}$$

The probability (g_i) that a healthy individual is from the ith subpopulation is

$$g_i = \frac{\gamma_i (1-p_i)}{\sum_{j=1}^n \gamma_j (1-p_j)} \quad , \tag{3}$$

which, for a rare disease (i.e., a small p_j in each population), is $\sim\gamma_i$.

Now, let q_i be the frequency of A in population i. If A is statistically independent of the disease (i.e., unassociated) *within* each subpopulation, it follows that

$$E(\hat{q}_d - \hat{q}_h) = \frac{\sum_{i=1}^n \gamma_i p_i q_i}{\sum_{j=1}^n \gamma_j p_j} - \frac{\sum_{i=1}^n \gamma_i (1-p_i) q_i}{\sum_{j=1}^n \gamma_j (1-p_j)} \quad ,$$

or, in the special case of just two populations,

$$E(\hat{q}_d - \hat{q}_h) = (p_1 - p_2)(q_1 - q_2) \times \left\{ \frac{\gamma_1 \gamma_2}{[\gamma_1 p_1 + \gamma_2 p_2][\gamma_1 (1-p_1) + \gamma_2 (1-p_2)]} \right\} \quad .$$

Hence, if both p_i and q_i vary across subpopulations, it is possible to have $E(\hat{q}_d - \hat{q}_h) \neq 0$. In that case, with a sufficiently large sample size, we can expect to find a (spurious) association between the candidate and the disease. Notice that the *reason* for the differences in disease frequencies among subpopulations—whether because of frequency differences at disease loci or differences in environmental factors—is not important in this model. A general feature of this type of model (implicit also in eq. [5] of Ewens and Spielman 1995) is that population structure is only an issue if the frequencies of both the disease *and* the marker alleles vary across subpopulations.

A second but less serious problem with population structure arises if the candidate locus is not in Hardy-Weinberg proportions, in which case the alleles within each individual are correlated. On its own, this would not cause the expectation of $\hat{q}_d - \hat{q}_h$ to be different from 0. However, in this case the test should use the two-allele genotypes (AA, AA^*, and A^*A^*) to obtain a proper χ^2 distribution under the null hypothesis. Use of two-allele genotypes would result in a 3×2 contingency table, and the test statistic could be computed as by Sokal and Rohlf (1995). We will not consider this issue further.

Relationship to relative risk.—Recall that f_i and g_i are the probabilities of a case and a control individual, respectively, being sampled from population i. Using the sampling model presented above, we can relate these probabilities to the relative risk of disease. We do this for the special case of two subpopulations; thus, f and g will denote the probabilities of sampled individuals coming from population 1. Using equations 2 and 3, we can write

$$\left(\frac{f}{1-f}\right)\left(\frac{1-g}{g}\right) = \left(\frac{p_1}{p_2}\right)\left(\frac{1-p_2}{1-p_1}\right) \quad ,$$

which, for a rare disease, is $\sim p_1/p_2$, the relative risk of disease in the two subpopulations. Expressing f in terms of g and the relative risk, RR, we obtain

$$f = \left(\frac{g}{1-g} RR\right)\left(1 + \frac{g}{1-g} RR\right)^{-1} \quad . \tag{4}$$

As we shall show, the probability of spurious associations increases rapidly with the relative risk.

Testing for Stratification

We now describe a procedure for testing whether the case and control samples are ethnically mismatched, using a series of unlinked markers. We will assume the sampling scheme described above.

Consider a study in which samples of unrelated cases and controls have been collected and in which it is unknown whether population structure should be a concern. To test whether the samples are ethnically matched, a set of L unlinked marker loci are typed in all the individuals in the original samples. Assume that the markers are chosen at random, so that it is improbable that any are tightly linked to disease loci. We want to know whether these markers indicate that the case and control samples are mismatched as a result of population subdivision. The null hypothesis is that the allele frequencies at each of the marker loci are the same in the case and control groups.

One way to test this is to construct contingency tables, as in table 1, that classify by case/control status and by allele or genotype at the marker locus. A χ^2 statistic can be computed for each of these tables; if biallelic markers are used, each test has the form of equation (1). Since we are interested in whether the loci show allele-frequency differences as a group, one natural test for stratification uses the sum of the test statistics from each locus. That is, we compute an overall test statistic, X_S^2, such that

$$X_s^2 = \sum_{i=1}^{L} X_i^2,$$

where X_i^2 is the χ^2 test statistic computed at the ith marker locus. Under the null hypothesis of no difference between the samples, X_S^2 is χ^2 distributed, with the number of df equal to the sum of the number of df of the individual loci.

How many loci?—It is clear that the power to detect stratification will depend on the number of loci used for the test. For this reason, it is important to choose a value for L large enough to ensure that the test has sufficient power to detect moderate stratification. For a given sampling scheme from a given population, let r_1 be the probability that a false association will be observed at a candidate locus (significant at the α level). Also, let r_2 be the probability of detecting stratification, with use of L unlinked marker loci, at the 5% level—say, conditional on the presence of an association at the candidate locus. The values of r_1 and r_2 depend on a number of factors that are not fully known to the investigator: the degree to which the cases and controls are mismatched, the degree of differentiation between subpopulations, and the precise details of the population history (which determine the distribution of allele frequencies across loci). The values of r_1 and r_2 also depend on L, α, and the sample sizes, which are, of course, determined by the investigator.

We are interested in the overall type I error rate for the experiment—that is, the probability of finding an association at the candidate locus but failing to detect stratification. Let R be the overall type I error rate for a given experiment; thus, $R = r_1(1 - r_2)$. Ideally, we would like to choose L in such a way that R is, at most, α, and is close to 0 if extreme stratification is present. In the case in which there is in fact no population subdivision, $R = .95\alpha$ and thus is mildly conservative. In the presence of stratification, r_1 and r_2 are larger than α and .05, respectively.

There are two types of random variation that affect the probability that a spurious association will be found and the probability that stratification will be detected. One is the sampling variance associated with the use of finite samples of cases and controls to estimate allele frequencies. The second is the difference, between populations, in allele frequencies, which may vary greatly from one locus to another. This variation is due to random sampling in the evolution of the populations. To model the second type of randomness, it is necessary to use an explicit evolutionary model of the populations. Detailed simulation results on the distribution of allele-frequency differences between populations have been published previously (Bowcock et al. 1991; Beaumont and Nichols 1996). Under certain conditions the differences can be modeled as the difference of two β-distributed random variables (Bowcock et al. 1991).

We have explored the properties of R in a series of simulations. These simulations assume that neither the candidate nor any of the other markers are linked to disease-susceptibility loci, but they permit spurious associations between the phenotype and the markers as the result of population structure.

Simulation Procedure

We assumed that case and control individuals were sampled from a structured population, according to the model described in "A Model of Stratified Sampling," with two subpopulations. We considered two models for the evolutionary divergence of the subpopulations. The results presented assume a standard model of population divergence without migration. That is, we assumed that a single ancestral population of effective population size (N_e) split, at some time (t) in the past, to produce two subpopulations, each of size N_e. The ancestral population was at mutation-drift equilibrium at the time of the split. We could adjust the degree of differentiation be-

tween subpopulations in this model by changing the amount of population divergence, tau, defined as t/N_e. We have also obtained similar results (not shown) with regard to choice of L, using a standard island-migration model (Wright 1951), despite the fact that the evolutionary-sampling process is quite different in these two cases.

In each replicate simulation, case individuals were sampled from subpopulation 1 with probability f and from subpopulation 2 with probability $1 - f$. Control individuals were sampled from subpopulations 1 and 2 with probabilities g and $1 - g$, respectively. Thus, the numbers of cases and controls from each subpopulation were binomially distributed random variables. There were equal numbers of cases and controls. Individuals were diploid.

Each replicate simulation included a candidate locus and a series of additional test loci. For each locus, we used a standard coalescent algorithm (Hudson 1990) to simulate a single ancestral genealogy, with the appropriate number of case and control chromosomes from each subpopulation. In the time between t and the present, chromosomes could only coalesce with other members of their own subpopulation; before time t, any pair of chromosomes could coalesce. Since we were simulating the case in which the disease phenotype is *independent* of the candidate and test loci, the choice of chromosomes joined at each coalescent event did not depend on phenotype. The number of cases and controls from each subpopulation was held constant within each replicate, but otherwise the genealogies were generated independently.

We considered two types of markers: microsatellites and biallelic markers. In the results shown, the candidate and the additional test loci were either all microsatellites or all biallelic markers. We modeled microsatellite mutation, using a pure stepwise unbiased mutation process without range constraints (Goldstein et al. 1995). In this kind of model, it is not necessary to specify the population size and mutation rate separately; instead, the mutation rate was specified in terms of $2N_e\mu$. The latter quantity is easily estimated for microsatellites, because it is equal to the expected variance in repeat scores under a stepwise model. We assumed a value of $2N_e\mu = 8.0$, which is typical of dinucleotide markers (Feldman et al. 1999). Variation at biallelic markers was generated by use of a low mutation rate (usually $2N_e\mu = 0.1$), and markers were selected only if the sample frequencies of both alleles was >.2. This threshold was chosen to approximate the likely characteristics of single-nucleotide-polymorphism surveys (Wang et al. 1998).

The statistical tests of association were performed as follows: the biallelic markers gave rise to 2 × 2 contingency tables, as in table 1. For each marker, we computed the χ^2 test statistic (see eq. [1]). The null distribution at the candidate locus was taken as being χ^2_1. To test for association, using the L unlinked marker loci, we computed the sum of the L-test statistics, obtained from equation (1). The null distribution of the sum was taken as being χ^2_L.

When the candidate locus was a microsatellite, we tested individual alleles for association with the phenotype. Alleles were tested only if their frequency exceeded a prespecified threshold (usually 10%). Each of the k alleles with frequencies above the threshold produced a 2 × 2 table, as in table 1. The candidate locus was considered to show a significant association if the largest test statistic for any allele exceeded a critical value. We obtained the critical value by approximating that the k-test statistics calculated were independent (each with a χ^2_1 distribution). We computed the critical values, using Sidák's multiplicative inequality (Rohlf and Sokal 1995, table E), with k tests and 1 df. In simulations (lowest lines in fig. 1*A* and *B*), we found that, because of the typically large numbers of alleles, this approximation was only very slightly conservative under the null model.

We used the L unlinked marker loci to test for stratification, as follows. Rare alleles (<10% frequency) were pooled either with each other or with the next rarest allele, if the pooled frequency was also <10%. For each locus, the k allelic classes that remained after pooling generated a 2 × k contingency table, from which a χ^2 test statistic was computed (Sokal and Rohlf 1995). The test statistics were summed across the L loci; the null distribution of the sum was determined to be χ^2, with $(\sum_{i=1}^{L} k_i) - L$ df.

To test the applicability of the model used to generate our simulated data, we also performed a series of simulation experiments, using genetic data from Jorde et al. (1995, 1997). These data included genotypes at 60 microsatellite loci from individuals from a number of human populations. There were 72 individuals of African origin (Biaka Pygmy, Mbuti Pygmy, Nguni, San, Sotho, or Tswana), and 120 individuals of European origin (British, Finnish, French, or Polish). Significant allele-frequency differences between the two groups were present at 75% of the loci.

We used the following procedure to simulate mismatched case-control samples from these populations. We drew (with replacement) two samples of 70 individuals from among the 192 individuals in the Jorde et al. data set. For one sample, the probability that each individual would be from the African group was .20, and for the other sample it was .20, .33, .50, and .66 in successive experiments (corresponding to relative risks of 1, 2, 4, and 8, respectively, with use of eq. [4]).

We then picked one locus at random, to be designated as a candidate locus. Using the selected individuals, plus their genotype data at the candidate locus, we performed

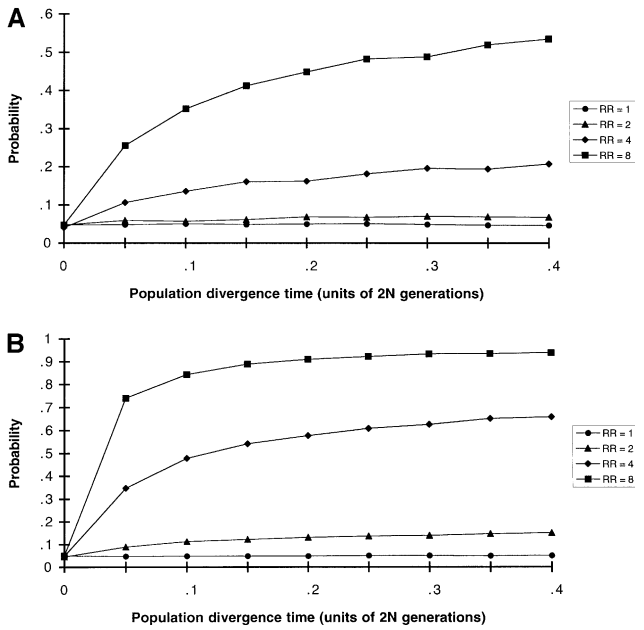

Figure 1 Probability that a spurious association will be detected at a microsatellite candidate locus (at the .05 significance level). The X axis (population divergence) is given in units of $2N_e$ generations; according to effective population size estimates in the work of Feldman et al. (1999), 1.0 corresponds to ~400,000 years, when a generation time of 20 years is assumed. Thus, the divergence between Africans and non-Africans is probably in the range .25–.40, and the divergence between non-African groups is probably <.20 (e.g., see Goldstein et al. 1995). Population relative risks in the range of 2–7 are typical of several common diseases, including non–insulin-dependent diabetes and hypertension (McKeigue 1997). The parameter values used were $g = .10$; $f = .10$, .18, .31, and .47, corresponding to relative risks of 1, 2, 4, and 8, respectively (see eq. [4]); and $2N\mu = 8.0$. *A*, Total sample size $m = 200$ individuals. *B*, Total sample size $m = 1{,}000$ individuals.

a test of association at the candidate locus, using the procedure described above. Then, if the candidate was significant at the 5% level, we picked an additional set of L marker loci (sampling from the remaining 59 loci with replacement). We used the additional markers to test for stratification between the two samples. We recorded the number of replicates in which the candidate locus was significant but in which the additional markers failed to indicate stratification.

Results

In figure 1*A* and *B*, we show the probability that a spurious association will be detected at a microsatellite candidate locus, as a function of several key parameters. The relative risk (RR) was computed from equation (4), under the assumption of a rare disease. When the relative risk is 1, with no sampling bias, the probability of detecting a spurious association is ~5%, regardless of the degree of population divergence. Otherwise, the probability that a spurious association will be detected increases with the relative risk, degree of population divergence, and sample size. For the smaller sample size (100 cases and 100 controls), the increase in false positives is negligible unless the relative risk in the two populations is >2. For the larger sample size (500 cases and 500 controls), the rate of false positives climbs more quickly for the small relative-risk groups. Results for biallelic candidate loci are similar (not shown).

Figure 2 shows the probability that stratification will be detected at the 5% level, with use of unlinked microsatellite markers. As expected, when tau = .0, so that there is no stratification, the probability of wrongfully inferring stratification is .05. Otherwise, when there is stratification, the power to detect stratification increases steadily with the number of marker loci used in the test and with the degree of stratification. Results obtained with other parameter values or biallelic markers are qualitatively similar.

As stated in "Testing for Stratification," what we would like to know is the probability R of obtaining a spurious association at the candidate locus and not detecting stratification, when an additional set of L unlinked marker loci is used. The probability of a spurious association increases with the degree of stratification (fig. 1), but so does the power to detect stratification (fig. 2).

In figure 3, we plot the overall type I error rate, R (defined as $r_1[1 - r_2]$), for a range of parameter values and for microsatellite and biallelic markers. Ideally, we would like to select the number of unlinked markers to ensure that the overall probability of believing a false association is no more than ~ .05. The degree of divergence in figure 3*A* and *C* is approximately as much as that observed between African and non-African populations; the degree of divergence in figure 3*B* and *D* is about as much as that observed between closely related

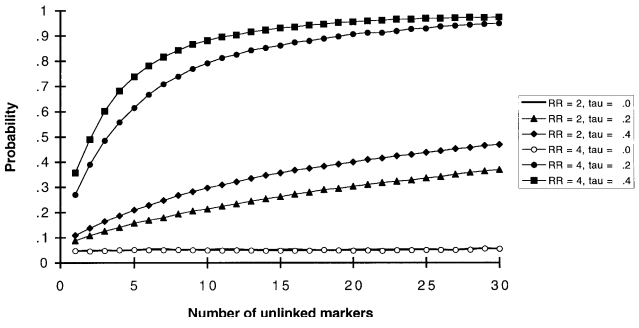

Figure 2 Probability that stratification will be detected, with use of unlinked microsatellite markers (at the .05 significance level). The parameter values are the same as in figure 1, with a total sample size of 200 individuals. The two lines at tau = .0 are on top of one another. As before, population divergence (tau = .2) corresponds to ~80,000 years.

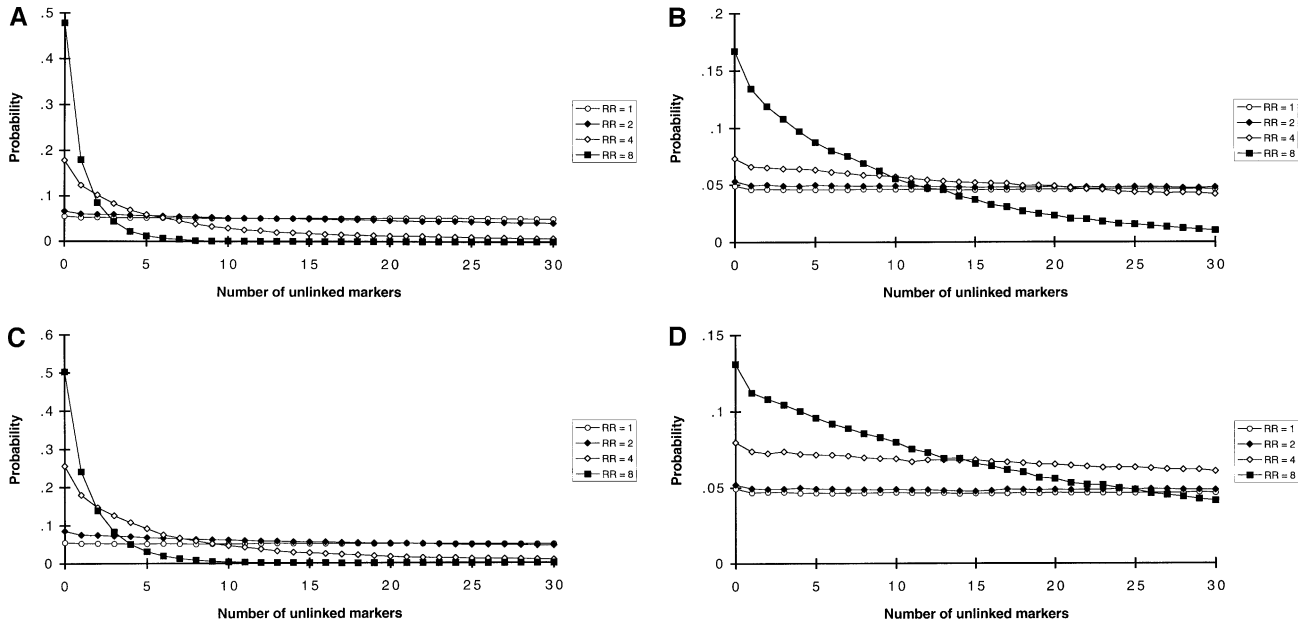

Figure 3 Probability that spurious association will be obtained at a candidate locus (at the .05 level) and *not* detecting stratification (at the .05 level), with use of unlinked marker loci. *A*, Microsatellite markers; tau = .25. *B*, Microsatellite markers; tau = .025. *C*, Biallelic markers; tau = .25. *D*, Biallelic markers; tau = .025. Recall that tau = .25 corresponds to a population divergence of ~100,000 years. Parameter values are as follows: sample size m = 200 individuals; g = .1; $2N\mu$ = 8.0 for microsatellites and .1 for biallelic markers.

ethnic groups but is more than that between most European populations (Cavalli-Sforza et al. 1994; Goldstein et al. 1995).

Ironically, when there is strong stratification, the overall risk of believing a false association is very low, even with small numbers of unlinked markers (fig. 3*A* and *C*). In that case, the power to detect stratification with a small number of markers is very high, which more than compensates for the high probability of detection of a spurious association at the candidate locus. More problematic is the situation with slight stratification (fig. 3*B* and *D*). In that case, the probability of a spurious association is only slightly greater than that in the unstratified case (.05); however, the power to detect stratification is also very low.

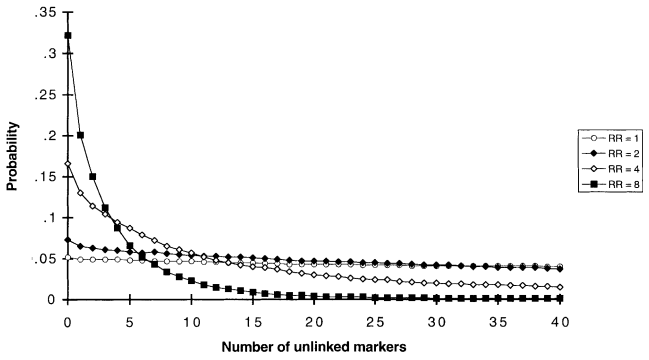

Figure 4 Probability that spurious association will be obtained at a candidate locus (at the .05 level) and *not* detecting stratification (at the .05 level), with use of unlinked marker loci estimated with microsatellite data collected by Jorde et al. (1995, 1997). The case and control groups were composed of different selections of Africans and Europeans, as described in the text.

We have run simulations, using a range of parameter values (i.e., varying the sample size, the degree of population divergence, and the relative risk). The upper lines in figure 3*B* and *D* are representative of the worst-case situations that we found. Those plots suggest that ~15–20 microsatellite loci are sufficient to test for stratification, which brings the overall type I error rate equal to the target rate of .05, even in the worst case. With biallelic markers, more loci are necessary. For the worst case, plotted in figure 3*D*, the overall type I error rate is .06, with 30 unlinked markers. This slight excess might be considered acceptable; if not, the requirement of a slightly more stringent cutoff at the candidate locus (i.e., a smaller α value) could be used as an alternative to the typing of more markers. For example, requiring significance at the .04 level at the candidate locus would reduce the overall probability of error to ~.05 in that worst case.

In figure 4 we show a summary of our results from the simulation experiments, which are based on the microsatellite data of Jorde et al. (1995, 1997). These results are consistent with the results that we obtained by

using simulated data (fig. 3) and support the estimate that 15–20 microsatellite loci are sufficient to bring the overall probability of error to no more than ~ 5%.

Discussion

Although the case-control–study design is often easier to implement than alternative approaches using the TDT, it is frequently criticized because of the potential for spurious associations resulting from population stratification. The case-control approach has generally been considered suspect, even if there is no prior reason to suspect the possibility of stratification. In response to this problem we show that population stratification can be detected by use of unlinked marker loci. In particular, strong stratification can be detected with high probability, with use of only a few markers.

In general, we have found that microsatellite markers provide more power to detect stratification than do biallelic markers. We recommend that case-control studies include ⩾15–20 unlinked microsatellites to test for stratification. If biallelic markers are used, more loci are needed to bring the overall type I error rate to ⩽5%, under the model considered. With 30 biallelic markers, the error rate can be as much as 6%. If this error rate is considered unacceptable, a slightly more stringent cutoff criterion could be used for the initial acceptance of an association at the candidate locus.

These guidelines were obtained by use of a pair of evolutionary models of population divergence (population splitting and island migration). We also presented similar results by using simulations done on the basis of human microsatellite data. A realistic concern is that exceptional loci might show unusually strong patterns of ethnic differentiation as the result of selection. When positive selection seems particularly likely (as in the major-histocompatibility-complex region), it would probably be wise to use more than the suggested number of markers to test for stratification. With the advent of genome screens, this issue should disappear, since the power to detect stratification with use of large numbers of markers will become very high.

We have focused on the situation in which there is no prior reason to suspect population structure, because case-control studies have often been criticized under such circumstances. However, when it is known that particular ethnic groups might have contributed to the sample (i.e., in admixed populations), it is appropriate to choose markers known to exhibit allele-frequency differences across the relevant populations, since this will clearly improve the power to detect stratification. In this case, fewer markers would be necessary (Shriver et al. 1997). Furthermore, it may be that particular microsatellite motifs tend to show more population differentiation than others (J. Pritchard, unpublished data). Even when there is little prior information about possible admixture, it might be sensible to use such markers preferentially in the test for stratification.

In the situation described above, we assumed that there was a candidate locus plus a series of markers used to test for stratification. In such a situation, there is a small, but non-0, probability that some of the test markers are actually linked to disease-susceptibility loci. Such an event would obviously increase the probability that stratification could be inferred, even when there is none (which makes the overall test conservative).

Strictly speaking, our results are concerned with the case in which only a single candidate locus is tested for association. However, these results should apply (approximately) in the case in which a small number of candidates are tested (with a Bonferroni correction). If many candidates are tested, then it may be that somewhat more loci are needed to test for stratification, to achieve the overall target error rate. In this case, a shortcut to reduce genotyping costs would be to include the candidate loci themselves in the test for stratification.

In particular, in a study testing large numbers of essentially random markers, such as in a genome screen, the frequency of markers that are in (nonspurious) linkage disequilibrium with disease genes is likely to be small. In that situation, it seems that a reasonable strategy is to use the genotyped markers themselves to test for stratification, so that no extra genotyping is necessary. In this case, given the large number of markers, the power to detect stratification will be very high indeed.

Acknowledgments

The authors thank L. Jorde for making his data available to them. They also thank D. Cox, M. Feldman, H. Jones, L. Lazzeroni, and N. Risch for helpful discussions and thank an anonymous reviewer for useful comments about the manuscript. J.K.P. was supported by a Howard Hughes predoctoral fellowship and by National Institutes of Health (NIH) grant GM19634; N.A.R. was supported by a National Defense Science and Engineering graduate fellowship. Part of this work was done while J.K.P. was visiting the Newton Institute, Cambridge. The research was supported in part by NIH grant GM28428 (to M. Feldman).

References

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B Biol Sci 263:1619–1626

Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am J Hum Genet 62:950–961

Bowcock AM, Hebert JM, Mountain JL, Kidd JR, Kidd KK, Cavalli-Sforza LL (1991) Drift, admixture, and selection in

human evolution: a study with DNA polymorphisms. Proc Natl Acad Sci USA 88:839–843

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton

Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57:455–464

Falk CT, Rubinstein P (1987) Haplotype relative risks: an easy way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Feldman MW, Kumm J, Pritchard JK (1999) Mutation and migration in models of microsatellite evolution. In: Goldstein D, Schlotterer C (eds) Microsatellites: evolution and applications. Oxford University Press, Oxford

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 92: 6723–6727

Hudson RR (1990) Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J (eds) Oxford surveys in evolutionary biology. Vol 7. Oxford University Press, Oxford, pp 1–44

Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, et al (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am J Hum Genet 57:523–538

Jorde, LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak R, Sung S, Kere J, et al (1997) Microsatellite diversity and the demographic history of modern humans. Proc Natl Acad Sci USA 94:3100–3103

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Lazzeroni LC, Lange K (1998) A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered 48:67–81

McKeigue PM (1997) Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. Am J Hum Genet 60:188–196

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Rohlf FJ, Sokal RR (1995) Statistical tables, 3d ed. WH Freeman, New York

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60: 957–964

Slatkin M (1991) Inbreeding coefficients and coalescence times. Genet Res 58:167–175

Sokal RR, Rohlf FJ (1995) Biometry, 3d ed. WH Freeman, New York

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am J Hum Genet 62:450–458

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–513

Thomson G, Robinson WP, Kuhner MK, Joe S (1989) HLA, insulin gene, and Gm associations with IDDM. Genet Epidemiol 6:155–160

Wang DG, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, Ghandour G, et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Wright S (1951) The genetical structure of populations. Ann Eugenics 15:323–354