Am. J. Hum. Genet. 63:241–251, 1998

Mapping Genes That Underlie Ethnic Differences in Disease Risk: Methods for Detecting Linkage in Admixed Populations, by Conditioning on Parental Admixture

Paul M. McKeigue

Epidemiology Unit, London School of Hygiene and Tropical Medicine, London

Summary

Genes that underlie ethnic differences in disease risk can be mapped in affected individuals of mixed descent if the ancestry of the alleles at each marker locus can be assigned to one of the two founding populations. Linkage can be detected by testing for association of the disease with the ancestry of alleles at the marker locus, by conditioning on the admixture (defined as the proportion of genes that have ancestry from the high-risk population) of both parents. With regard to exploiting the effects of admixture, this test is more flexible and powerful than the transmission-disequilibrium test. Under the assumption of a multiplicative model, the statistical power for a given sample size depends only on parental admixture and the risk ratio *r* between populations that is generated by the locus. The most informative families are those in which mean parental admixture is .2–.7 and in which admixture is similar in both parents. The number of markers required for a genome search depends on the number of generations since admixture and on the information content for ancestry (*f*) of the markers, defined as a function of allele frequencies in the two founding populations. Simulations using a hidden Markov model suggest that, when admixture has occurred 2–10 generations earlier, a multipoint analysis using 2,000 biallelic markers, with *f* values of 30%, can extract 70%–90% of the ancestry information for each locus. Sets of such markers could be selected from libraries of single-nucleotide polymorphisms, when these become available.

Received December 1, 1997; accepted for publication May 12, 1998; electronically published June 19, 1998.

Address for correspondence and reprints: Dr. Paul M. McKeigue, Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom. E-mail: p.mckeigue@lshtm.ac.uk

0002-9297/98/6301-0034$02.00

Introduction

When there has been recent admixture between two populations that, for genetic reasons, have different disease risks and when the ancestry of the alleles at marker loci can be assigned to one of these two founding populations, the gametic disequilibrium generated by admixture can be exploited to map the genes that underlie these ethnic differences in disease risk (Chakraborty and Weiss 1988). If a marker locus is linked to a locus where genetic variation underlies a difference, in disease risk, between the two founding populations, the proportion of alleles at the marker locus that have ancestry from the high-risk population will be higher in affected individuals than that expected by chance (McKeigue 1997).

If admixture (defined as the proportion of an individual's genome that has ancestry from the high-risk population) varies among individuals in the admixed population, the risk of disease will vary with admixture. The frequency of alleles that have ancestry from the high-risk population therefore will be higher in affected than in unaffected individuals sampled from the population of mixed descent. For example, type 2 diabetes in populations of mixed European/Native American descent is associated with markers of Native American ancestry, such as *GM* haplotypes, that presumably are unlinked to the disease (Knowler et al. 1988). To exploit the effects of admixture, to map genes that underlie ethnic differences in disease risk, we require a statistical test that eliminates the association in the ancestry of alleles at unlinked loci that is generated by variation in the overall admixture in individuals in the population.

In an initial exploration of the admixture approach (McKeigue 1997), I suggested use of the transmission-disequilibrium test (Ewens and Spielman 1995) to test for excess transmission of alleles that have ancestry from the high-risk population to affected offspring of parents who have inherited, at the marker locus, one allele from each of the two founding populations. This article explores the properties of an alternative test for linkage in admixed populations, by conditioning on the overall admixture of each parent. This test detects association only

in the presence of linkage and is shown to be more powerful and flexible than the transmission-disequilibrium test, in exploiting the linkage disequilibrium generated by admixture. The sample size required for detection of linkage by this approach is examined for various genetic models and values of parental admixture. Statistical methods by which ancestry of the alleles at each locus can be estimated, by combining marker data at all loci for each family in a multipoint analysis, are outlined. The ancestry information (f) conveyed by a marker is derived as a function of allele frequencies in the two founding populations, and the number of markers required for a genome search is estimated by simulation.

Test for Association Conditional on Parental Admixture

Comparison with the Transmission-Disequilibrium Test

We consider a population in which admixture has occurred between a low-risk population X and a high-risk population Y, and we denote alleles that have ancestry from population Y as "Y by descent." We define the admixture of an individual as the probability that an allele chosen at random from a locus chosen at random from that individual is Y by descent. Individuals whose admixture is $\frac{1}{2}$ are referred to as "equally admixed," and individuals who have one allele, at a locus, derived from each of the two founding populations are referred to as "heterozygous for ancestry" at that locus.

Using the transmission-disequilibrium test, we can test for excess transmission of alleles Y by descent, from parents heterozygous for ancestry at the marker locus to affected offspring. When used in this manner, the transmission-disequilibrium test is a test for association conditional on the ancestry of parental alleles at the marker locus. One limitation of a test conditioning on parental ancestry at the marker locus is that it requires both parents to be available for genotyping, unless parental genotypes can be inferred from the genotypes of unaffected sibs. This limits the possibilities for the study of late-onset conditions such as non–insulin-dependent diabetes. A more fundamental disadvantage is that conditioning on parental ancestry at the marker locus does not fully exploit the information generated by admixture, except when parents are from the F_1 generation and therefore are heterozygous for ancestry at all loci.

This can be seen by considering a simple example. Suppose that we study an autosomal recessive Mendelian trait for which the frequency of the high-risk allele is 0 in population X and 1 in population Y. Suppose that mixed unions occur between individuals in populations X and Y and that the offspring of these mixed unions form an endogamous subpopulation within which random mating produces successive generations equivalent to the (F_2, F_3, F_4, …) generations produced in an experimental cross between inbred strains. In a sample of n affected individuals from the F_3 and subsequent generations, all $2n$ alleles at the trait locus will be Y by descent. Half the $2n$ parents of these affected individuals will be heterozygous for ancestry at the trait locus, and half will have two alleles Y by descent at the trait locus (the frequencies of the three possible mating types XY/XY, XY/YY, and YY/YY are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively). We could use the transmission-disequilibrium test to examine the n alleles transmitted, from parents heterozygous for ancestry, to affected offspring and to test the proportion of these n alleles that are Y by descent, for departure from the proportion $\frac{1}{2}$, expected under the null hypothesis. Alternatively, we could examine all $2n$ alleles at the trait locus in affected offspring and test the proportion of these $2n$ alleles that are Y by descent, for departure from the proportion $\frac{1}{2}$, expected from the overall admixture of their parents. Because the transmission-disequilibrium test is restricted to transmissions from heterozygous parents, it uses only half the linkage information that is present in the F_3 and subsequent generations.

A general expression for the proportion Π of alleles at the disease locus that have ancestry from the high-risk population, in affected offspring, is derived in Appendix A. For the affected offspring of equally admixed parents, this proportion is the same whether we examine only the alleles transmitted from parents heterozygous, for ancestry, at the disease locus (eq. [A1]) or all alleles transmitted to these offspring (eq. [A2]). For the F_3 and subsequent generations in which parental genotypes are in Hardy-Weinberg equilibrium, the probability that a parent is heterozygous for ancestry at the disease locus, given that the offspring are affected, is $\frac{1}{2}$. Since only half the parents of affected individuals are heterozygous for ancestry at the disease locus, a transmission-disequilibrium test will require twice as many families as a test conditioning on parental admixture, to detect departure from the null hypothesis that $\Pi = \frac{1}{2}$. Comparisons show that, over a wide range of values of parental admixture, the required sample size for a transmission-disequilibrium test is approximately twice as large as that for a test conditioning on parental admixture.

Test Conditional on Parental Admixture as a Test for Linkage

In testing for association between a trait and the ancestry of alleles conditional on parental admixture, we are testing for gametic disequilibrium in ancestry, between alleles at the trait locus and alleles at the marker locus, that is independent of parental admixture. We can argue by induction that if there has been no selection of alleles at these loci, since admixture, then this is a specific

test for linkage between the marker locus and the trait locus.

Let M_i be the admixture of the ith parent in a population, and consider two loci, A and B, in a gamete produced by this parent. Let A_{Yi} be the event that the gamete carries an allele Y by descent at locus A, and let B_{Yi} be the event that the gamete carries an allele Y by descent at locus B. Suppose that the events A_{Yi} and B_{Yi} are independent, given that parental admixture has the value M_i. This is equivalent to the assumption that

$$P(B_{Yi}|A_{Yi}) = P(B_{Yi}) = M_i \ . \qquad (1)$$

If equation (1) holds, the ith parent will produce gametes of ancestry A_YB_Y, A_YB_X, A_XB_Y, and A_XB_X, in the proportions M_i^2, $M_i(1 - M_i)$, $M_i(1 - M_i)$, and $(1 - M_i)^2$, respectively. Consider the offspring of two parents with admixture M_1 and M_2. The 16 possible parental mating types are shown in table 1, together with the proportions in which offspring of each mating type produce gametes of ancestry A_YB_Y and A_YB_X, if the loci are unlinked. Multiplication of the frequency of each mating type by the proportion in which each type of gamete is produced by offspring of this mating type and summation of these product terms over mating types show that offspring of these parents produce gametes of ancestry A_YB_Y and A_YB_X in the proportions M^2 and $M(1 - M)$, respectively, where $M = \frac{1}{2}(M_1 + M_2)$. Therefore, the probability that a gamete produced by these offspring carries an allele Y by descent at locus B is equal to M, whether or not the gamete carries an allele Y by descent at locus A. Thus, if equation (1) holds for the gametes produced by all parents in one generation and if loci A and B are unlinked, then it will hold for the next generation also. In gametes from either of the two founding populations, disequilibrium in the ancestry of alleles that is independent of parental admixture cannot occur: specifying admixture as 0 (gametes from population X) or 1 (gametes from population Y) specifies ancestry at all loci. It follows that equation (1) holds for all generations that are of mixed descent, if loci A and B are unlinked, whatever the history of admixture. It follows that, for two unlinked loci, disequilibrium in the ancestry of alleles that is independent of parental admixture cannot arise except by chance. If we condition on parental admixture, a test for association between a trait and the ancestry of alleles at a marker locus is therefore a specific test for linkage.

Relation of Required Sample Size to Population Risk Ratio and Admixture

Sample Size in an Equally Admixed Population

Assuming equal admixture in the parental generation, we can use equation (A3) in Appendix A to examine how much the required sample size is likely to depend on the underlying genetic model. Let r be the population risk ratio generated by the locus—that is, the ratio of disease risk in population Y to disease risk in population X. Equation (A3) can be simplified to an expression for Π that depends only on r, for the following four simple models:

1. For the multiplicative model $f_2 = \gamma f_1 = \gamma^2 f_0$, where γ is the genotypic risk ratio and f_i is the penetrance of the genotype with i copies of the high-risk allele (Risch and Merikangas 1996), $\Pi = \frac{1}{2} + (\sqrt{r} - 1)/(2\sqrt{r} + 2)$;
2. For the additive model $f_2 - f_1 = f_1 - f_0$, $\Pi = \frac{1}{2} + (r - 1)/(4r + 4)$;
3. For the recessive model $f_1 = f_0$, with the low-risk allele fixed (allele frequency equal to 1) in the low-risk population, $\Pi = \frac{1}{2} + (r - 1)/(2r + 6)$; and
4. For the dominant model $f_2 = f_1$, with the high-risk allele fixed in the high-risk population, $\Pi = \frac{1}{2} + (r - 1)/(6r + 2)$.

Under the null hypothesis, Π has a value of $\frac{1}{2}$. For a given value of r, Π is furthest from $\frac{1}{2}$ for a recessive model

Table 1

Frequencies with Which Gametes of Each Type Are Produced by Offspring of Parents with Admixture M_1 and M_2, for Two Unlinked Loci

Gamete [Frequency] from Parent 1	Proportions in Which Offspring Produce Gametes of Types A_YB_Y, A_YB_X, If Loci A and B Are Unlinked, When Gamete [Frequency] from Parent 2 Is			
	A_XB_X $[(1 - M_2)^2]$	A_XB_Y $[M_2(1 - M_2)]$	A_YB_X $[M_2(1 - M_2)]$	A_YB_Y $[M_2^2]$
A_XB_X $[1 - M_1)^2]$	0, 0	0, 0	0, $\frac{1}{2}$	$\frac{1}{4}$, $\frac{1}{4}$
A_XB_Y $[M_1(1 - M_1)]$	0, 0	0, 0	$\frac{1}{4}$, $\frac{1}{4}$	$\frac{1}{2}$, 0
A_YB_X $[M_1(1 - M_1)]$	0, $\frac{1}{2}$	$\frac{1}{4}$, $\frac{1}{4}$	0, 1	$\frac{1}{2}$, $\frac{1}{2}$
A_YB_Y $[M_1^2]$	$\frac{1}{4}$, $\frac{1}{4}$	$\frac{1}{2}$, 0	$\frac{1}{2}$, $\frac{1}{2}$	1, 0

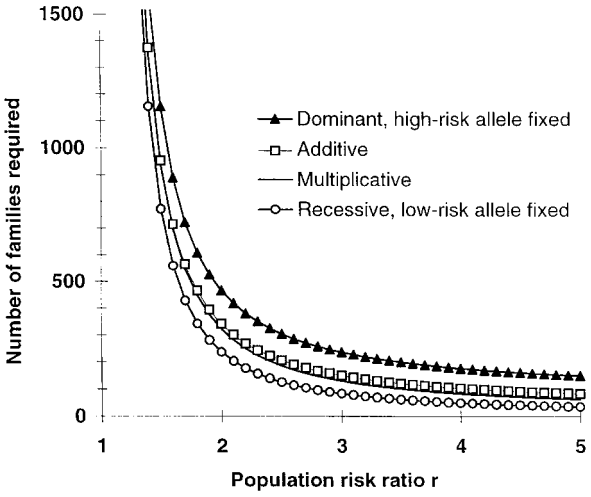

Figure 1 Relation of required sample size to *r*, under four possible genetic models: additive, multiplicative, recessive with low-risk allele fixed in low-risk population, and dominant with high-risk allele fixed in high-risk population. Sample sizes are for 90% power to detect locus, at $P < .001$.

with the low-risk allele fixed in the low-risk population X and closest to $\frac{1}{2}$ for a dominant model with the high-risk allele fixed in the high-risk population Y. Since the sample size required to detect departure from the null hypothesis depends on the deviation of Π from its null value, these two extreme models thus define the smallest and largest sample sizes required to detect linkage, for a given value of *r*. Figure 1 shows the relation of required sample size (calculated for 90% power to detect at $P < .001$, by use of a one-sided test) to *r*, for these four models. When $r \geqslant 2$, a sample of a few hundred families has adequate statistical power, whatever the underlying genetic model. When $1.5 < r \leqslant 2$, between 300 and 1,000 families are required under a multiplicative model, but, for a given value of *r*, the sample size required under a dominant or recessive model does not vary by more than approximately one-third that required under a multiplicative model. In contrast, the statistical power of an affected-sib-pair design is critically dependent on the allele frequency and the genotypic risk ratio (Risch and Merikangas 1996): only when the frequency of the high-risk allele is <.25 and the genotypic risk ratio is $\geqslant 3$ (corresponding to a sibling recurrence-risk ratio $\geqslant 1.36$) does the affected-sib-pair design have 90% power to detect a locus, at $P < .001$, with <500 families.

Effect of Unequal Admixture on Required Sample Size

The above analyses are based on the assumption of an equally admixed population. Even when admixture is unequal, under a multiplicative model the statistical power of the admixture design can be shown to depend only on parental admixture and the *r* that is generated by the locus. We write $P(Y_j)$ for the probability that when one allele is chosen at random from each parent, from a locus chosen at random, *j* of these two alleles are Y by descent. The probabilities $P(Y_0)$, $P(Y_1)$, and $P(Y_2)$ then are determined by admixture M_1 and M_2 of parents 1 and 2, respectively. Substituting $f_2 = \gamma f_1 = \gamma^2 f_0$ into equation (A2) in Appendix A, we obtain an expression for the proportion Π of alleles at the disease locus that have ancestry from the high-risk population, in affected offspring, that depends only on *r* and the admixture of both parents:

$$\Pi = \frac{\frac{1}{2}P(Y_1)\sqrt{r} + P(Y_2)r}{P(Y_0) + P(Y_1)\sqrt{r} + P(Y_2)r} .$$

When parents belong to a homogeneous population of mixed descent or when mating within the admixed population is highly assortative with respect to admixture (e.g., if socioeconomic status is closely related to admixture), admixture generally will be similar in both parents. Figure 2 shows the relation of required sample size, for 90% power to detect at $P < .001$, to admixture *M*, assumed (for illustration) to be the same in all parents. The sample size is smallest when *M* is ~.4 but remains close to its minimum value for any values of *M* between ~.2 and ~.7.

When admixture differs between the two parents, the probability $P(Y_1)$, in relation to $P(Y_0)$ and $P(Y_2)$, will be larger than that expected for Hardy-Weinberg frequencies, and the sample size required to detect departure from the null hypothesis $r = 1$ will be larger than that required when admixture is the same in both parents.

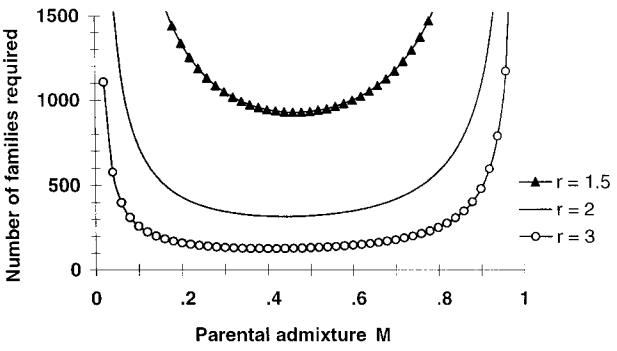

Figure 2 Relation of required sample size to *M* (assumed to be the same in all parents) and *r* generated by the locus, under the assumption of a multiplicative model. Sample sizes are for 90% power to detect locus, at $P < .001$.

Figure 3 shows the effect on sample size if admixture differs between the two parents, on the basis of a multiplicative model and $r = 2$. We assume, for the purpose of illustration, that the sample consists entirely of families with the same pair of values for M_1 and M_2. If mean parental admixture is close to $\frac{1}{2}$, the required sample size is ~15% larger when $|M_1 - M_2| = .25$ and ~75% larger when $|M_1 - M_2| = .5$, compared with the sample size required when both parents are equally admixed.

When mean parental admixture varies outside the range .3–.6 and when there are large differences between the admixture of the two parents, statistical power for a given sample size is reduced markedly. Thus, for a sample consisting entirely of families for which $M_1 = .5$ and $M_2 = 0$ or 1 (equivalent to a backcross between an equally admixed individual and an individual from one of the two founding populations), the required sample size is approximately three times larger than that required when both parents are equally admixed.

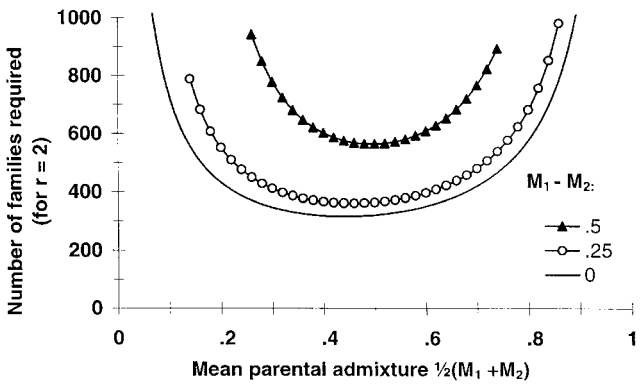

Figure 3 Relation of required sample size to mean parental admixture, under the assumptions of the same pair of parental admixture values M_1 and M_2 in each family, a multiplicative model, and $r = 2$. Sample sizes are for 90% power to detect locus, at $P < .001$.

Multipoint Statistical Methods for Assignment of Ancestry at Each Locus

The above analyses demonstrate the theoretical limit of the statistical power of the admixture design to detect linkage, if for each affected individual the ancestry of the alleles at each locus can be assigned accurately as 0, 1, or 2 alleles Y by descent. Even if we choose marker loci at which an allele that is absent in one of the two founding populations is common in the other, we usually will not be able to assign the ancestry of alleles at a locus by typing a single marker. Thus, if we rely on analyzing markers one locus at a time, the statistical power of an admixture study will not be anywhere near its theoretical limit (Kaplan et al. 1998). With a multipoint analysis, however, information from closely spaced marker loci can be combined so as to assign ancestry at each locus accurately, even though no single marker is fully informative for ancestry.

The underlying principle of such a multipoint analysis is simple. We first choose a set of marker polymorphisms for which there are large allele-frequency differences between populations X and Y and space these markers at a much higher density than the density of transitions of ancestry (between X by descent and Y by descent) on chromosomes of individuals of mixed descent. If we type these markers in an affected individual, together with those in the individual's parents, sibs, or offspring, we can assign haplotypes and reconstruct the sequence of marker alleles on each chromosome. Over any short interval, the haplotype in an individual of mixed descent will consist mainly of alleles that are more common in one of the two founding populations than in the other. By combining the information from these marker alleles, we can reduce the uncertainty with which the ancestry of the alleles at each locus is assigned as 0, 1, or 2 alleles Y by descent. We now examine possible approaches to developing a multipoint method that simultaneously assigns haplotypes, combines information from closely spaced markers to estimate ancestry of each allele at each marker locus, and tests for linkage with the trait under study.

One method would be to extend the approach that has been developed, by Lander and Green (1987) and Kruglyak and Lander (1995), for classic linkage studies. In this approach, each meiosis that gives rise to a nonoriginal (an individual with one or more parents in the pedigree) is modeled as a Markov process in which each locus corresponds to one step in the chain. The chain state at each locus is defined as 0 or 1 according to whether the paternally derived or the maternally derived marker allele is transmitted. For each locus, the chain states for all meioses in the pedigree are combined into an inheritance vector. Standard methods for hidden Markov models (MacDonald and Zucchini 1997) then are applied to calculate the probability distribution of the inheritance vectors at each locus, conditional on the marker data for all loci simultaneously. If the entire history of admixture for the individual under study is known so that a pedigree can be constructed in which each original is from one of the founding populations, it is, in principle, straightforward to extend this hidden Markov model to estimate the ancestry of each allele at each locus, conditional on the marker data. When admixture dates back more than three generations, however, it is unlikely that the pedigree that traces the ancestry of the individual under study back to originals in the founding populations will be known.

A variant of the hidden Markov model approach, in

which the history of admixture for the individual under study is unknown, would be to construct a simpler pedigree, based on the individual under study plus any parents, sibs, or offspring who have been typed. We then could model the transitions of ancestry on each set of chromosomes inherited from a parent of an original, as a two-state Markov process (with unknown generator matrix) on a continuous axis. Standard methods for hidden Markov models then could be used to estimate the probability distribution at each locus, for the ancestry of the alleles transmitted to the affected individual. The problem with this approach is that the transitions of ancestry on chromosomes inherited from a parent of mixed descent do not necessarily follow a first-order Markov process. When densely spaced markers with high information content are used, a first-order Markov model still may be a satisfactory approximation for purposes of estimating ancestry at each locus. Alternatively, it may be possible to fit a higher-order hidden Markov model, for which methods have been described elsewhere (MacDonald and Zucchini 1997).

An alternative approach, which would overcome the limitations of hidden Markov models, would be to use Markov chain simulation (Thompson 1994; Gelman and Rubin 1996) to estimate the probability distribution of ancestry at each locus, conditional on all marker data and on any available information about the history of admixture. If all unknown quantities are treated as random variables and if prior distributions for these variables are specified, the Gibbs sampler or other sampling algorithms can be used to construct a Markov chain that converges on the joint posterior distribution of these variables, given the observed data. Thus, when the history of admixture is unknown, we can define random variables to specify a pedigree that traces the ancestry of the individual under study back to originals in one of the two founding populations. The values of each coordinate of the inheritance vector—conditional on the transmissions at adjacent loci, the ancestry of the originals, and the observed marker data—then could be updated by the Gibbs sampler. To test for linkage, one could define a score as the vector of the observed number minus the expected number of alleles Y by descent and could obtain the expectation of this score statistic and its variance (the information matrix) by averaging over the posterior distribution of the missing data, given the observed data (Little and Rubin 1987, pp. 127–140). When the history of admixture is known, the estimates of ancestry obtained by Markov chain simulation would be identical to those obtained analytically from a hidden Markov model, since both methods are based on the same underlying statistical model.

Although in principle Markov chain simulation can correctly estimate the distribution of states of ancestry at each locus, conditional on the observed marker data, when the Gibbs sampler is applied to linkage analyses, it usually is necessary to modify the sampling algorithm to ensure that the Markov chain adequately explores the underlying probability space (Sobel and Lange 1993; Lin et al. 1994). Before any multipoint method is used in practice, extensive simulation would be required, to test the robustness of the method when allele frequencies or prior distributions are misspecified.

Marker Information Content for Ancestry

The ancestry information conveyed by a marker polymorphism can be measured by the extent to which typing an allele at the marker locus reduces our uncertainty about the ancestry of the allele. Since the ancestry information conveyed by a single marker varies according to the prior probabilities of ancestry from each founding population, it has been suggested that markers should be selected according to the admixture of the population under study (Kaplan et al. 1998). In a multipoint analysis, however, the dependence of information extracted by markers on admixture is not necessarily the same as that in a single-point analysis. If we score ancestry at a locus as 0 when the allele is X by descent and 1 when the allele is Y by descent, the variance of ancestry of this allele in an equally admixed population is $\frac{1}{4}$ when no information about allele type is available and $\pi(1-\pi)$ when the allele has been typed, where π is the posterior probability that the allele is Y by descent, given the allele type. Thus, the variance of ancestry is reduced by a proportion equal to $1-4\pi(1-\pi)$. We define the marker information content for ancestry between two populations as the expected proportion f by which variance of ancestry at a marker locus is reduced when an allele at this locus is typed, for a population with equal admixture from these two populations. For a biallelic marker, this proportion is a function of the frequencies u_X and u_Y of allele 1 in populations X and Y, respectively:

$$f = \frac{(u_X - u_Y)^2}{4\bar{u}(1-\bar{u})} \text{ , where } \bar{u} = \tfrac{1}{2}(u_X + u_Y) \text{ .} \qquad (2)$$

Thus, an f value of 30% corresponds to $\bar{u} = .23$ and $|u_X - u_Y| = .46$ or to $\bar{u} = .5$ and $|u_X - u_Y| = .55$. f is equal to the standardized variance of allele frequencies originally defined by Wahlund (1928). For two populations, f represents the proportion by which heterozygosity at the locus is reduced as a result of division into two populations of equal size that have different allele frequencies, in relation to the heterozygosity of a population formed by pooling these two populations. The average f value of markers is closely related to the genetic distance between the two populations, defined as the fixation index F_{ST}, which is an estimate of the average proportion by which heterozygosity has been reduced

since the two populations diverged from a common ancestral population (Wright 1951; Cavalli-Sforza et al. 1994). If F_{ST} distances are calculated with a correction for sampling error (Reynolds et al. 1983), the mean f value of markers is slightly more than half the F_{ST} distance.

Distribution of Marker Information Content

We can estimate, from published surveys of allele frequencies in different populations, what proportion of biallelic markers are likely to have f values above any given cutoff level, between any given pair of founding populations. Large differences, in allele frequencies, between two populations can result either from drift or from disruptive selection, as, for instance, when a marker is in linkage disequilibrium with a locus that influences susceptibility to malaria. The distribution of F_{ST} values for marker polymorphisms has been found to be more skewed than would be expected under drift alone (Bowcock et al. 1991), suggesting that at least one-fifth of markers have been under disruptive selection. Thus, this means that markers that have large f values (>20%) between two populations can be found easily, even when the mean f value between these two populations is <10%.

In two recent surveys of allele frequencies for restriction-site polymorphisms in unadmixed Europeans and Africans (Jorde et al. 1995; Poloni et al. 1995), the mean f value between the two populations was ~8%. Although these f values may be inflated by sampling error in the estimates of allele frequencies, in these two studies most markers were typed on samples of at least 150 chromosomes in each population, so that this bias is likely to be small. Of the markers typed, 8% (8/95) had f values, between Europeans and Africans, that were >20%, and the mean f value for markers that were above this cutoff was ~30%. A few markers with $f > 70\%$, between Europeans and Africans, are known, such as the *FY* and *GM* loci (Cavalli-Sforza et al. 1994). In future, the main sources of large numbers of biallelic markers are likely to be the libraries of single-nucleotide polymorphisms (SNPs) that are now being assembled (National Human Genome Research Institute 1998). If the distribution of f values for SNPs is similar to that for biallelic restriction-site polymorphisms, we can estimate that, to identify 2,000 markers that have an average f value of 30%, between Europeans and an African population, it will be necessary to screen a library of $\geqslant$30,000 SNPs.

The F_{ST} distance from Europeans to west Africans is ~.15 (Cavalli-Sforza et al. 1994). The F_{ST} distances from Europeans to Native Australians (.15), from Europeans to Pacific Islanders (.13), and from Europeans to Native Americans (.11) are not much smaller than the F_{ST} distance from Europeans to west Africans. Thus, it may be possible to identify sets of markers that have high f values between these pairs of populations also, by screening libraries of SNPs. Identification of markers that have high f values between Europeans and Indians, for which the F_{ST} distance is only ~.05 (Cavalli-Sforza et al. 1994), is likely to be more difficult.

Information Extracted by Markers, by Use of Simulation with a Hidden Markov Model

The above analyses suggest that it will be possible to identify a set of 2,000 markers that have average f values of 30%, between any two groups that are separated by an F_{ST} distance comparable to that between Europeans and west Africans. We can use simulation to examine the ancestry information that is extracted by a multipoint analysis using such markers. For the purpose of this exercise, we assume that haplotypes have been reconstructed accurately. The problem of assigning ancestry at each locus then is reduced to that of combining the information from a sequence of marker alleles on a set of chromosomes inherited from the same parent, so as to estimate the ancestry of each allele. For an equally admixed population within which random mating generates successive generations equivalent to the (F_2, F_3, F_4, ...) generations in an experimental cross, Appendix B shows that the transitions of ancestry on chromosomes can be approximated by a Markov process so that a hidden Markov model can be applied.

A two-state Markov chain representing ancestry transitions between marker loci on a chromosome was simulated for a locus flanked on either side by 10 evenly spaced markers. Spacing between markers was set to three alternative values: 1 cM, 1.5 cM, or 2 cM. Transition probabilities were set to represent chromosomes from generations F_2–F_{10}. Marker allele frequencies in populations X and Y were chosen to give an f value of 30% for each marker. At each locus, an allele type was generated randomly from the probability distribution specified by the ancestry and the ancestry-specific frequency of the allele. The probability π that the allele at the locus under examination is Y by descent, conditional on all 21 markers, was calculated from a hidden Markov model as in Appendix B. Information extraction is estimated as the mean value of $1 - 4\pi(1 - \pi)$, on the basis of 500 simulations: this is the proportionate reduction, in variance of ancestry, that results from typing the markers.

The results of this simulation are shown in figure 4. With marker spacing of 1.5 cM, information extraction at the locus falls from ~90% in chromosomes from generation F_2 to ~70% in chromosomes from generation F_{10}. In practice, information extraction would be lower than this, because there would be some uncertainty in

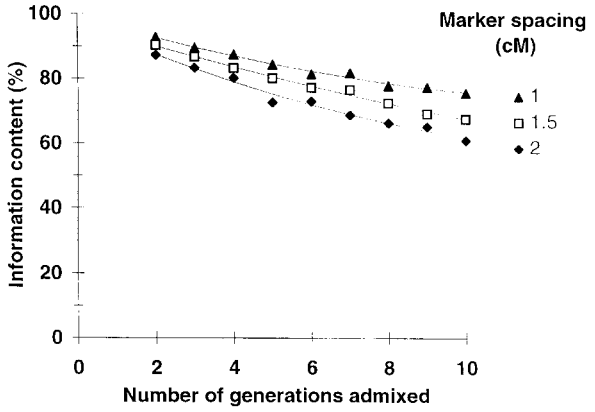

Figure 4 Information content extracted by a hidden Markov model estimating ancestry at a marker locus flanked by 10 evenly spaced markers on each side, by use of markers with an *f* value of 30%, on chromosomes of individuals from generations F_2–F_{10}.

the reconstruction of haplotypes and because, when the history of admixture is unknown, parameters for the model have to be estimated from the data. The information extracted does not depend on the marker allele frequencies as long as these are chosen to give an *f* value of 30% for each marker. Thus, results determined by use of this simple model suggest that markers with average *f* values of 30%, spaced at intervals of 1.5 cM (equivalent to ~2,000 markers if the length of the genome is 3,000 cM), will be adequate for a genome search in populations in which admixture dates back $\leqslant 10$ generations.

Discussion

The admixture design is analogous to linkage analysis of an experimental cross between two inbred strains, generalized to situations in which the founding populations are not inbred and the ancestry of individuals of mixed descent is not under experimental control or is not even known. Since linkage analyses of experimental crosses between strains are generally more powerful than linkage analyses within strains, for the detection of quantitative-trait loci (Kruglyak and Lander 1995), it is not surprising that similar advantages in statistical power apply to the admixture design, compared with conventional allele-sharing designs for detection of genes of modest effect. The only new technological development required for the admixture design is the availability of a library of $\geqslant 30{,}000$ biallelic polymorphisms, together with automated methods for typing them. Assembly of libraries of SNPs is now in process (National Human Genome Research Institute 1998).

On the basis of the results in this paper, it is possible to outline an experimental strategy for exploiting admixture to map genes underlying ethnic differences in disease risk, for populations in which admixture has occurred within the past few hundred years. The first step is to establish whether ethnic differences in disease risk are likely to have a genetic basis. The strongest evidence for this will come from population-based association studies that show a relationship between risk of disease and proportionate admixture from the high-risk population. A set of markers suitable for genome-wide assignment of ancestry is identified by screening libraries of SNPs or other biallelic polymorphisms. The information content for ancestry of each marker is estimated, and a set of $\geqslant 2{,}000$ markers that have information content >20% is chosen. When representative samples of the populations from which founders originated are not available, a multipoint method could be used to reestimate allele frequencies in the two founding populations, from genotypes of families of mixed descent. This also would correct the allele frequencies, for drift or selection since admixture.

To map genes underlying the ethnic difference in disease risk, affected individuals of mixed descent are typed, together with their parents, if available. Where parents are not available, sibs or offspring of affected individuals are typed. Information about the history of admixture is obtained when possible, and the investigator attempts to collect families in which both parents of the affected individual have similar values of admixture, with a mean value of .2–.7. A multipoint method, based on one of the approaches outlined previously, is used to assign haplotypes, estimate ancestry at each locus, and test for linkage. The approach can be generalized easily to quantitative traits, by sampling individuals whose trait values vary from the value expected from their admixture.

Acknowledgments

I am grateful to Tim Aitman for his comments on the manuscript and to Steve Bennett, David Clayton, and Ian White for their comments on the statistical analysis.

Appendix A

Ancestry at Disease Locus in Affected Individuals

We consider a disease locus with two alleles: a high-risk allele D_1 and a low-risk allele D_2. Let f_0, f_1, and f_2 be the penetrances of genotypes D_2D_2, D_1D_2, and D_1D_1, respectively. Let the frequencies of alleles D_1, D_2 be p_X, q_X,

and p_Y, q_Y in populations X and Y, respectively. Suppose that an individual is taken at random from a population in which there has been recent admixture between populations X and Y and that the ancestry of alleles at the disease locus can be assigned as 0, 1, or 2 alleles Y by descent. Let F be the event that the individual is affected, and let G_i be the event that the individual's genotype at the disease locus has i copies of the D_1 allele. We define Y_j as the event that, when one allele is chosen at random from each parent, at a locus chosen at random, j of these two alleles are Y by descent. The probabilities $P(Y_0)$, $P(Y_1)$, and $P(Y_2)$ thus are determined by the admixture of each parent.

Conditioning on Parental Ancestry at the Marker Locus

Let H be the event that the parent is heterozygous for ancestry at the disease locus, T the event that an allele Y by descent is transmitted, and S_j the event that j alleles at the disease locus in the offspring are Y by descent. Let M be the probability that the allele transmitted from the other parent is Y by descent. Then, the probabilities $P(S_0|T)$, $P(S_1|T)$, and $P(S_2|T)$ are 0, $1 - M$, and M, respectively. Similarly, the probabilities $P(S_0|H)$, $P(S_1|H)$, and $P(S_2|H)$ are $\frac{1}{2}(1 - M)$, $\frac{1}{2}$, and $\frac{1}{2}M$, respectively.

For a test conditional on parental ancestry at the disease locus (or a nearby marker locus), we require the probability Π' of transmission of an allele Y by descent, at the disease locus from a parent heterozygous for ancestry at this locus, given that the individual is affected:

$$\Pi' = P[T|(H \text{ and } F)] = \frac{P(F|T)P(T|H)}{P(F|H)} .$$

We have $P(T|H) = \frac{1}{2}$, $P(F|T) = \sum_j \sum_i f_i P(G_i|S_j)P(S_j|T)$, and $P(F|H) = \sum_j \sum_i f_i P(G_i|S_j)P(S_j|H)$. Hence,

$$\Pi' = \frac{\frac{1}{2}\sum_j \sum_i f_i P(G_i \mid S_j)P(S_j \mid T)}{\sum_j \sum_i f_i P(G_i \mid S_j)P(S_j \mid H)}$$

$$= \frac{\frac{1}{2}(1 - M)[q_X q_Y f_0 + (q_X p_Y + p_X q_Y)f_1 + p_X p_Y f_2] + \frac{1}{2}M(q_Y^2 f_0 + 2p_Y q_Y f_1 + p_Y^2 f_2)}{\frac{1}{2}(1 - M)(q_X^2 f_0 + 2p_X q_X f_1 + p_X^2 f_2) + \frac{1}{2}[q_X q_Y f_0 + (q_X p_Y + p_X q_Y)f_1 + p_X p_Y f_2] + \frac{1}{2}M(q_Y^2 f_0 + 2p_Y q_Y f_1 + p_Y^2 f_2)} . \quad \text{(A1)}$$

Conditioning on Parental Admixture

We require the probability Π that an allele at the disease locus, transmitted to an affected individual, is Y by descent, given event F and the probabilities $P(Y_0)$, $P(Y_1)$, and $P(Y_2)$:

$$\Pi = \frac{1}{2}P(Y_1|F) + P(Y_2|F) = \frac{\frac{1}{2}P(F|Y_1)P(Y_1) + P(F|Y_2)P(Y_2)}{P(F)} .$$

We have $P(F|Y_j) = \sum_i f_i P(G_i|Y_j)$ and $P(F) = \sum_j P(F|Y_j)P(Y_j)$. Hence,

$$\Pi = \frac{\frac{1}{2}P(Y_1)\sum_i f_i P(G_i \mid Y_1) + P(Y_2)\sum_i f_i P(G_i \mid Y_2)}{\sum_j \sum_i f_i P(G_i \mid Y_j)P(Y_j)}$$

$$= \frac{\frac{1}{2}P(Y_1)[q_X q_Y f_0 + (q_X p_Y + p_X q_Y)f_1 + p_X p_Y f_2] + P(Y_2)(q_Y^2 f_0 + 2p_Y q_Y f_1 + p_Y^2 f_2)}{P(Y_0)(q_X^2 f_0 + 2p_X q_X f_1 + p_X^2 f_2) + P(Y_1)[q_X q_Y f_0 + (q_X p_Y + p_X q_Y)f_1 + p_X p_Y f_2] + P(Y_2)(q_Y^2 f_0 + 2p_Y q_Y f_1 + p_Y^2 f_2)} . \quad \text{(A2)}$$

In a population in which all parents are equally admixed, the probability M is equal to $\frac{1}{2}$, and the probabilities $P(Y_0)$, $P(Y_1)$, and $P(Y_2)$ are equal to $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. Substituting these values into equations (2) and (A1) yields the same expression for Π' and Π:

$$\Pi' = \Pi = \frac{1}{2}\left[1 + \frac{(q_Y^2 - q_X^2)f_0 + 2(p_Y q_Y - p_X q_X)f_1 + (p_Y^2 - p_X^2)f_2}{(q_X + q_Y)^2 f_0 + 2(p_X + p_Y)(q_X + q_Y)f_1 + (p_X + p_Y)^2 f_2}\right]. \quad \text{(A3)}$$

The r generated by the locus is given by

$$r = \frac{p_Y^2 f_2 + 2p_Y q_Y f_1 + q_Y^2 f_0}{p_X^2 f_2 + 2p_X q_X f_1 + q_X^2 f_0}.$$

For the four genetic models described in the text, expressions for Π in terms of r can be obtained by substitution into the above equations.

Appendix B

Fitting a Hidden Markov Model to Transitions of Ancestry on Chromosomes from the F_n Generation of Mixed Descent

We consider a population in which the offspring of mixed unions form an endogamous subpopulation, within which random mating produces successive generations equivalent to the (F_2, F_3, F_4, ...) generations produced by an experimental cross between inbred strains. In a set of chromosomes from generation F_n, the coefficient Δ_n of disequilibrium of ancestry between two loci separated by recombination fraction θ (equivalent to a map distance of x morgans) is $\Delta_1 = \frac{1}{4}$, and, for $n \geqslant 2$,

$$\Delta_n = \tfrac{1}{4}(1 - 2\theta)(1 - \theta)^{n-2}$$

(as shown elsewhere [McKeigue 1997]), which can be rewritten as

$$\Delta_n = \tfrac{1}{4}e^{-2x}\left(\tfrac{1}{2} + \tfrac{1}{2}e^{-2x}\right)^{n-2} = \tfrac{1}{4}e^{-nx}\cosh^{n-2}x$$

$$\simeq \tfrac{1}{4}e^{-nx}, \text{ for small values of } x.$$

The matrix $\mathbf{T}$ of transition probabilities is then given by

$$\mathbf{T} = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{-nx} & \frac{1}{2} - \frac{1}{2}e^{-nx} \\ \frac{1}{2} - \frac{1}{2}e^{-nx} & \frac{1}{2} + \frac{1}{2}e^{-nx} \end{bmatrix}.$$

This is the transition matrix of a two-state Markov process on a continuous axis, in which the exponential distribution parameters for the lengths of segments of each state have the value $\frac{1}{2}n$, and the stationary distribution δ is $[\frac{1}{2}\ \frac{1}{2}]$. Although this simple model holds only for an equally admixed population within which all mating occurs between successive nonoverlapping generations of mixed descent, we can use it to estimate the number of markers required for a genome search, as a function of marker spacing, marker information content for ancestry, and number of generations of admixture. Suppose that marker alleles have been typed at a sequence of m loci on a single chromosome. For all values of i from 1 to m, we calculate for the ith locus (1) a 2×2 diagonal matrix $\mathbf{Q}_i$, in which the diagonal elements are the probabilities of the observed marker allele at locus i, given each of the two possible states of ancestry at this locus, and (2) the 2×2 transition matrix $\mathbf{T}_i$ between locus i and locus $i + 1$. To calculate the probability distribution of ancestry at locus t, we calculate row vectors $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ as $\boldsymbol{\alpha}_t = \delta\mathbf{Q}_1\mathbf{T}_1\mathbf{Q}_2\mathbf{T}_2 \ldots \mathbf{Q}_t$ and $\boldsymbol{\beta}_t' = \mathbf{T}_t\mathbf{Q}_{t+1}\mathbf{T}_{t+1}\mathbf{Q}_{t+2}\mathbf{T}_{t+2} \ldots \mathbf{T}_{m-1}\mathbf{Q}_m 1'$. The expression $[\text{diag}(\boldsymbol{\alpha}_t)\text{diag}(\boldsymbol{\beta}_t)]/\boldsymbol{\alpha}_t\boldsymbol{\beta}_t'$ gives a 2×2 di-

agonal matrix in which the diagonal elements are the probabilities of each possible state of ancestry at locus *t*, conditional on marker data at all *m* loci (MacDonald and Zucchini 1997).

References

Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, 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, NJ

Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci USA 85: 9119–9123

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

Gelman A, Rubin DB (1996) Markov chain Monte Carlo methods in biostatistics. Stat Methods Med Res 5:339–355

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

Kaplan NL, Martin ER, Morris RW, Weir BS (1998) Marker selection for the transmission/disequilibrium test, in recently admixed populations. Am J Hum Genet 62:703–712

Knowler WC, Williams RC, Pettitt DJ, Steinberg AG (1988) $Gm^{3;5,13,14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am J Hum Genet 43:520–526

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84: 2363–2367

Lin S, Thompson E, Wijsman E (1994) An algorithm for Monte Carlo estimation of genotype probabilities on complex pedigrees. Ann Hum Genet 58:343–357

Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York

MacDonald IL, Zucchini W (1997) Hidden Markov and other models for discrete-valued time series. Chapman & Hall, London

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

National Human Genome Research Institute (1998) Methods for discovering and scoring single nucleotide polymorphisms. RFA HG-98-001, National Institutes of Health, Bethesda, MD

Poloni ES, Excoffier L, Mountain JL, Langaney A, Cavalli-Sforza LL (1995) Nuclear DNA polymorphism in a Mandenka population from Senegal: comparison with eight other human populations. Ann Hum Genet 59:43–61

Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the co-ancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–779

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

Sobel E, Lange K (1993) Metropolis sampling in pedigree analysis. Stat Methods Med Res 2:263–282

Thompson EA (1994) Monte Carlo likelihood in the genetic mapping of complex traits. Philos Trans R Soc Lond B Biol Sci 344:345–350

Wahlund S (1928) Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas 11:65–106

Wright S (1951) The genetical structure of populations. Ann Eugenics 15:322–354