Am. J. Hum. Genet. 74:1136–1153, 2004

Linkage Analysis of a Complex Disease through Use of Admixed Populations

Xiaofeng Zhu,[1] Richard S. Cooper,[1] and Robert C. Elston[2]

[1]Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, Maywood, IL; and [2]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland

Linkage disequilibrium arising from the recent admixture of genetically distinct populations can be potentially useful in mapping genes for complex diseases. McKeigue has proposed a method that conditions on parental admixture to detect linkage. We show that this method tests for linkage only under specific assumptions, such as equal admixture in the parental generation and admixture that occurs in a single generation. In practice, these assumptions are unlikely to hold for natural populations, resulting in an inflation of the type I error rate when testing for linkage by this method. In this article, we generalize McKeigue's approach of testing for linkage to allow two different admixture models: (1) intermixture admixture and (2) continuous gene flow. We calculate the sample size required for a genomewide search by this method under different disease models: multiplicative, additive, recessive, and dominant. Our results show that the sample size required to obtain 90% power to detect a putative mutant allele at a genomewide significance level of 5% can usually be achieved in practice if informative markers are available at a density of 2 cM.

Introduction

Genomewide linkage analysis has been widely used as a first step in attempts to positionally clone genes influencing complex traits. However, the power of such studies is often low if the effect of the gene on the trait is modest (Risch and Merikangas 1996). Association studies using the information generated by recent population admixture (or admixture mapping) may provide an alternative to family-based linkage analysis, because significant linkage disequilibrium (LD) between two loci can persist over several centimorgans in admixed populations (Lautenberger et al. 2000; Pfaff et al. 2001; Risch 1992; Stephens et al. 1994). In the United States, the African American and Mexican American populations could be considered potential candidate populations for admixture mapping, because their DNA segments are derived from African/European and Native American/European ancestry, respectively. The LD created by admixture in African Americans can be detectable for as much as 30 cM after as many as 20 generations (Stephens et al. 1994; Lautenberger et al. 2000; Pfaff et al. 2001). Because of the presence of long-range LD in an admixed population, the density of markers required for a whole-genome search can be reduced in comparison with what is required for an association study.

Rife (1954) was the first to suggest the use of an admixed population to detect linkage, specifically in the context of continuous traits, using correlation analysis. Since then, several methods have been proposed to make use of the disequilibrium arising from recent admixture of two populations with differing trait and marker allele frequencies in the mapping of genes for complex binary diseases (Chakraborty and Weiss 1988; Risch 1992; Stephens et al. 1994; Thomson 1995; McKeigue 1997, 1998; Kaplan et al. 1998; Zheng and Elston 1999; Halder and Shriver 2003). Chakraborty and Weiss (1988) assumed that the disequilibrium between alleles at the marker and disease loci could be measured directly, which is an assumption more suitable for Mendelian diseases (McKeigue 1997). They suggested using a likelihood-ratio test based on the history of admixture, which can be difficult to model, because we usually lack detailed information on the history of the populations. Because gametic disequilibrium generated by admixture persists for a few generations even for unlinked loci, Stephens et al. (1994) suggested excluding from the analysis those individuals whose ancestry changed within the past few generations preceding the study, to avoid inflating the probability of a false-positive error. Adopting this procedure can exclude many informative samples and can make study recruitment more difficult, however. To avoid these difficulties, McKeigue (1997) suggested using the

Received September 2, 2003; accepted for publication March 12, 2004; electronically published May 6, 2004.

Address for correspondence and reprints: Dr. Xiaofeng Zhu, Department of Preventive Medicine and Epidemiology, Loyola University Medical Center, 2160 South First Avenue, Maywood, IL 60153. E-mail: xzhu1@lumc.edu

0002-9297/2004/7406-0008$15.00

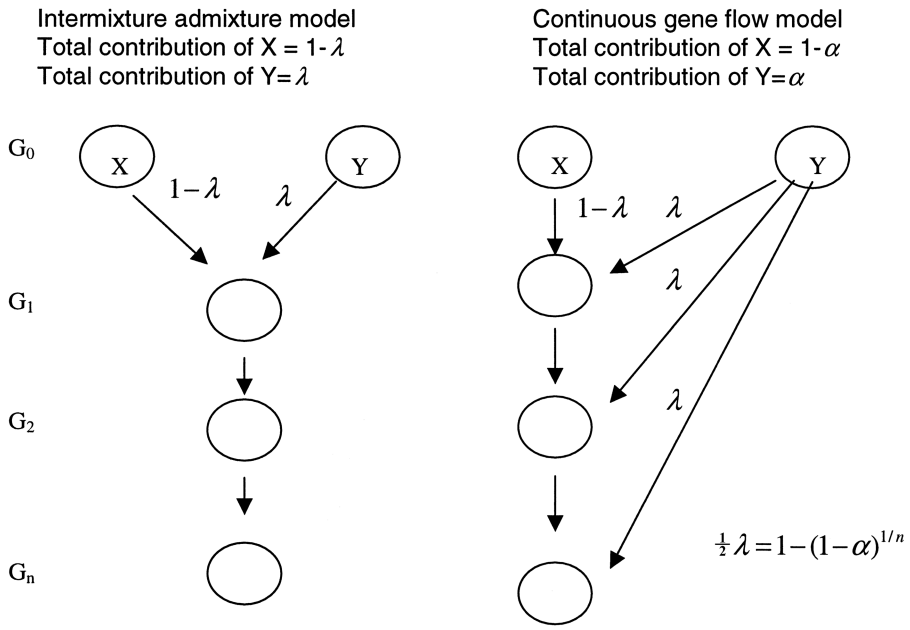

Figure 1 Two admixture models used in the calculation of ancestral probability. *Left panel,* IA model. *Right panel,* CGF model.

transmission/disequilibrium test (TDT) (Spielman et al. 1993) to test for excess transmission to affected children of alleles with different frequencies in the parental populations. Because the degree of association between two loci depends on the allele frequency differences in the founding populations, Kaplan et al. (1998) further suggested that the alleles within a locus should be grouped according to the information they provide regarding the parental and admixed populations, thus improving power. Zheng and Elston (1999) introduced a multipoint method based on the TDT method, for mapping the trait locus position through use of admixed populations. McKeigue (1998) proposed an intriguing method to test for gametic disequilibrium between alleles at two loci conditional on the parental admixture. He demonstrated that testing for association between loci conditional on parental admixture is a direct test for linkage. On the basis of this idea, a score test adopting a Bayesian approach was proposed (McKeigue et al. 2000). However, the claim that testing for "association of ancestry conditional on parental admixture" is a test for linkage requires the assumptions that the admixed population is an equal mixture of the two parental populations and that the admixture occurs in a single generation, followed by recombination and drift, with no further genetic contribution from either parental population. In practice, these assumptions may not be valid. For example, Pfaff et al. (2001) demonstrated that the African American population has more likely experienced a continuous gene flow (CGF) pattern of admixture. In an abstract, Risch (1992) proposed an interesting idea that involved sampling singleton cases for association analysis in admixed populations, but no further details were provided. In this article, we theoretically calculate the probability of the ancestral origin of a marker locus in an affected individual under two admixture models: (1) intermixture admixture (IA) and (2) CGF. We then propose a case-only method of testing for linkage that can, with the appropriate population, be more powerful than allele-sharing linkage analysis. We investigate the power to test for linkage under different disease modes of inheritance: multiplicative, additive, recessive, and dominant.

Models

IA Model

Suppose we have an admixed population C (such as African Americans or Mexican Americans) resulting from two parental populations, X and Y (for example, African and European or American Indian and European). At the first generation of the admixture process, we let $1 - \lambda$ be the proportion of offspring of two persons from population X and let λ be the proportion of offspring with one parent from X and the other from Y. In the following generations, mating takes place at random, regardless of the ancestry of the mates (fig. 1, *left panel*). We also assume there is no mutation or selection. This model has been designated "the IA model" (Long 1991).

Table 1

Conditional Probabilities of Haplotype Gametes (*h*) Produced Given Parental Two-Locus Genotypes (*g*)

	Probability Given Genotype									
	$\frac{M_X}{D}\|\frac{M_X}{D}$	$\frac{M_X}{D}\|\frac{M_X}{d}$	$\frac{M_X}{d}\|\frac{M_X}{d}$	$\frac{M_X}{D}\|\frac{M_Y}{D}$	$\frac{M_X}{D}\|\frac{M_Y}{d}$	$\frac{M_X}{d}\|\frac{M_Y}{D}$	$\frac{M_X}{d}\|\frac{M_Y}{d}$	$\frac{M_Y}{D}\|\frac{M_Y}{D}$	$\frac{M_Y}{D}\|\frac{M_Y}{d}$	$\frac{M_Y}{d}\|\frac{M_Y}{d}$
Haplotype	(g_1)	(g_2)	(g_3)	(g_4)	(g_5)	(g_6)	(g_7)	(g_8)	(g_9)	(g_{10})
M_XD (h_1)	1	1/2	0	1/2	$(1-\theta)/2$	$\theta/2$	0	0	0	0
M_Xd (h_2)	0	1/2	1	0	$\theta/2$	$(1-\theta)/2$	1/2	0	0	0
M_YD (h_3)	0	0	0	1/2	$\theta/2$	$(1-\theta)/2$	0	1	1/2	0
M_Yd (h_4)	0	0	0	0	$(1-\theta)/2$	$\theta/2$	1/2	0	1/2	1

Let D and d represent the two alleles at a disease locus, with allele frequencies p_X, q_X in population X, and p_Y, q_Y in population Y, respectively. Let M be a marker locus, and let θ be the recombination fraction between the marker locus and the disease locus. On the assumption of the IA model, at the first generation there are 7 of 10 possible genotypes (g) present, shown in table 1: M_XD/M_XD, M_XD/M_Xd, M_Xd/M_Xd, M_XD/M_YD, M_XD/M_Yd, M_Xd/M_YD, and M_Xd/M_Yd, with probabilities $(1-\lambda)p_X^2$, $2(1-\lambda)p_Xq_X$, $(1-\lambda)q_X^2$, λp_Xp_Y, λp_Xq_Y, λq_Xp_Y, and λq_Xq_Y, respectively, where the slash mark indicates haplotype phase. Here, the subscripts "X" and "Y" represent the population of origin of an allele. Table 1 presents the frequencies of the gametes produced by all possible parental genotypes.

Let $h_1^{(1)}$, $h_2^{(1)}$, $h_3^{(1)}$, and $h_4^{(1)}$ be the haplotype frequencies produced in generation 1. Then we have

$$h_1^{(1)} = (1-\lambda)p_X + \frac{\lambda}{2}[p_X - (p_X - p_Y)\theta] ,$$

$$h_2^{(1)} = (1-\lambda)q_X + \frac{\lambda}{2}[q_X - (q_X - q_Y)\theta] ,$$

$$h_3^{(1)} = \frac{\lambda}{2}[p_Y + (p_X - p_Y)\theta] ,$$

and

$$h_4^{(1)} = \frac{\lambda}{2}[q_Y + (q_X - q_Y)\theta] .$$

Because we assume that mating is at random, the probabilities of the marker allele being from populations X and Y are $1-(\lambda/2)$ and $\lambda/2$, respectively, in all generations. The frequencies of D and d remain $[1-(\lambda/2)]p_X + (\lambda/2)p_Y$ and $[1-(\lambda/2)]q_X + (\lambda/2)q_Y$, respectively, in the following generations. The haplotype frequencies in the nth generation can be easily calculated using the iterative formulas:

$$\begin{aligned}
h_1^{(n)} &= (1-\theta)h_1^{(n-1)} + \theta\left(1-\frac{\lambda}{2}\right)\left[\left(1-\frac{\lambda}{2}\right)p_X + \frac{\lambda}{2}p_Y\right] ,\\
h_2^{(n)} &= (1-\theta)h_2^{(n-1)} + \theta\left(1-\frac{\lambda}{2}\right)\left[\left(1-\frac{\lambda}{2}\right)q_X + \frac{\lambda}{2}q_Y\right] ,\\
h_3^{(n)} &= (1-\theta)h_3^{(n-1)} + \frac{\lambda\theta}{2}\left[\left(1-\frac{\lambda}{2}\right)p_X + \frac{\lambda}{2}p_Y\right] ,\\
\text{and } h_4^{(n)} &= (1-\theta)h_4^{(n-1)} + \frac{\lambda\theta}{2}\left[\left(1-\frac{\lambda}{2}\right)q_X + \frac{\lambda}{2}q_Y\right] .
\end{aligned} \tag{1}$$

Let f_0, f_1, and f_2 be the penetrances of the disease genotypes DD, Dd, and dd, respectively. We further assume that the penetrances are the same in both parental populations. Let us also use the term "X by descent," as defined by McKeigue (1998), to denote an allele having ancestry from X. Let $\Pi^{(n)}(\theta)$ be the proportion of marker alleles X by descent (the "ancestral probability" from population X) among affected individuals at the nth generation since admixture occurred. Then we can write

$$\begin{aligned}
\Pi^{(n)}(\theta) = \frac{1}{P(\text{affected})}\sum_i & P(\text{affected} \mid g_i)\\
& \times P(\text{an allele is X by descent} \mid g_i)P(g_i) ,
\end{aligned}$$

where g_i is one of the 10 possible genotypes shown in table 1, and P(affected) represents the disease prevalence at the nth generation in the admixed population. At the nth generation, Hardy-Weinberg equilibrium proportions hold, and D and d have frequencies

$(1/2)[(2-\lambda)p_X + \lambda p_Y]$ and $(1/2)[(2-\lambda)q_X + \lambda q_Y]$, respectively. Thus,

$$P(\text{affected}) = \frac{1}{4}f_2[(2-\lambda)p_X + \lambda p_Y]^2 + \frac{1}{2}f_1[(2-\lambda)p_X + \lambda p_Y] \times [(2-\lambda)q_X + \lambda q_Y] + \frac{1}{4}f_0[(2-\lambda)q_X + \lambda q_Y]^2 .$$

From table 1 and the formulas in (1), after some algebra we obtain

$$\Pi^{(n)}(\theta) = \frac{2-\lambda}{2} + \frac{(2-\lambda-2\theta)\lambda(1-\theta)^{n-2}}{8} \times \frac{\xi}{P(\text{affected})} , \quad (2)$$

where

$$\xi = \{(f_2 - f_1)[(2-\lambda)p_X + \lambda p_Y] + (f_1 - f_0)[(2-\lambda)q_X + \lambda q_Y]\}(p_X - p_Y) .$$

Let

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

represent the relative risk ratio of parental population X to Y. We then obtain, letting γ be the genotypic risk ratio:

1. for the multiplicative model $f_2 = \gamma f_1 = \gamma^2 f_0$,

$$\Pi^{(n)}(\theta) = \frac{2-\lambda}{2} + \frac{(2-\lambda-2\theta)\lambda(1-\theta)^{n-2}}{2} \times \frac{\sqrt{r}-1}{(2-\lambda)\sqrt{r}+\lambda} ;$$

2. for the additive model $f_2 - f_1 = f_1 - f_0$,

$$\Pi^{(n)}(\theta) = \frac{2-\lambda}{2} + \frac{(2-\lambda-2\theta)\lambda(1-\theta)^{n-2}}{4} \times \frac{r-1}{(2-\lambda)r+\lambda} ;$$

3. for the recessive model $f_1 = f_0$, with $p_Y = 0$,

$$\Pi^{(n)}(\theta) = \frac{2-\lambda}{2} + \frac{(2-\lambda-2\theta)\lambda(1-\theta)^{n-2}}{2} \times \frac{(2-\lambda)(r-1)}{(2-\lambda)^2(r-1)+4} ;$$

and

4. for the dominant model $f_2 = f_1$, with $p_X = 1$,

$$\Pi^{(n)}(\theta) = \frac{2-\lambda}{2} + \frac{(2-\lambda-2\theta)\lambda(1-\theta)^{n-2}}{2} \times \frac{\lambda(r-1)}{(1-r)\lambda^2+4r} .$$

If population C is equally admixed by A and B and the marker and disease loci are at the same position, we have $\theta = 0$ and $\lambda = 1$. In this case, all of the formulas in the article by McKeigue (1998, p. 243) follow.

It can be seen that $p_X = p_Y$ is equivalent to $r = 1$ when $f_2 \geqslant f_1 > f_0$. From equation (2) and the results for models 1–4, we see that testing the null hypothesis $H_0{:}\Pi^{(n)}(\theta) = (2-\lambda)/2$ is equivalent to testing $\theta = 1 - (\lambda/2)$ or $r = 1$. Note that, at all generations, $(2-\lambda)/2$ is the ancestral probability that a marker allele comes from parental population X. When $\lambda = 1$, the null hypothesis becomes $H_0{:}\Pi^{(n)}(\theta) = 1/2$, which is equivalent to testing $\theta = 1/2$ or $r = 1$. McKeigue (1998) proposed a method he called "conditioning on parental admixture," to test for linkage between a marker and a disease locus by testing the null hypothesis $H_0{:}\Pi^{(n)}(\theta) = 1/2$ (McKeigue 1998, p. 243). He subsequently let the null hypothesis $H_0{:}\Pi^{(n)}(\theta) = 1/2$ be equivalent to the null hypothesis $H_0{:}r = 1$ (p. 244). However, $H_0{:}\Pi^{(n)}(\theta) = 1/2$ is equivalent to the joint test of $\theta = 1/2$ and $r = 1$, provided there is an equal admixture rate from each parental population. In practice, we rarely observe a population with an equal admixture rate, and then only by following a very special sampling scheme. It should be pointed out that $r = 1$ should not be considered as a null hypothesis and that $r \neq 1$ is a necessary condition for $H_0{:}\Pi^{(n)}(\theta) = 1/2$ to be equivalent to $\theta = 1/2$.

From equation (2), $\Pi^{(n)}(\theta)$ is always a strict monotonic function of the recombination fraction θ between a marker locus and a disease locus when $r \neq 1$, independent of the marker allele frequencies. After the first two generations, $\Pi^{(n)}(\theta)$ asymptotically tends to the expected ancestral probability from the parental population X, at a rate of $(1-\theta)^{n-2}$. $\Pi^{(n)}(\theta)$ achieves its maximum or minimum value when $\theta = 0$. Whether it is a maximum or minimum value depends on which ancestral population has a higher frequency of the D allele. Thus, to estimate

the location of the disease locus, we can find the position maximizing or minimizing $\Pi^{(n)}(\theta)$, provided we are able to estimate $\Pi^{(n)}(\theta)$.

CGF Model

In the previous section, we discussed a model in which admixture occurs in the first generation alone, followed by recombination and drift, with no further genetic contribution from either parental population. We now study the CGF model, in which admixture occurs at a steady but reduced rate in each generation (fig. 1, *right panel*) (Long 1991; Pfaff et al. 2001). This model is more likely than the IA model to mimic observed experience with the African American population. Neither mutation nor selection are considered here.

In appendix A, we show how, under the CGF model, $\Pi^{(n)}(\theta)$ can be expressed as a function of p_X, p_Y, θ, the generation n, and the penetrance functions. Furthermore, we show that it is a monotonic function of θ. Thus, $\Pi^{(n)}(\theta)$ has properties under the CGF model similar to those under the IA model. Therefore, to estimate the position of the disease locus, we again find the position maximizing or minimizing $\Pi^{(n)}(\theta)$. Using the formula for $\Pi^{(n)}(\theta)$ in appendix A and the model assumptions, we obtain

1. for the multiplicative model,

$$\Pi^{(n)}(\theta) = \frac{(1-\frac{\lambda}{2})^n\left[(1-\lambda)(1-\frac{\lambda}{2})^n(\sqrt{r}-1)+1-\frac{\lambda}{2}\right]\left[1+(1-\frac{\lambda}{2})^{-1}c_{22}(\sqrt{r}-1)\right]}{\left[(1-\lambda)(1-\frac{\lambda}{2})^n(\sqrt{r}-1)+1\right]\left[(1-\frac{\lambda}{2})^n(\sqrt{r}-1)+1\right]};$$

2. for the additive model,

$$\Pi^{(n)}(\theta) = \frac{(1-\frac{\lambda}{2})^n\left[0.5(1-\lambda)(1-\frac{\lambda}{2})^n(r-1)+0.5c_{22}(r-1)+1-\frac{\lambda}{2}\right]}{0.5(2-\lambda)(1-\frac{\lambda}{2})^n(r-1)+1};$$

3. for the recessive model $f_1 = f_0$, with $p_Y = 0$,

$$\Pi^{(n)}(\theta) = \frac{(1-\frac{\lambda}{2})^n\left[(1-\lambda)(1-\frac{\lambda}{2})^{n-1}c_{22}(r-1)+1-\frac{\lambda}{2}\right]}{(1-\lambda)(1-\frac{\lambda}{2})^{2n}(r-1)+1};$$

and

4. for the dominant model $f_2 = f_1$, with $p_X = 1$,

$$\Pi^{(n)}(\theta) = \frac{(1-\frac{\lambda}{2})^n\left[\left[1-(1-\lambda)(1-\frac{\lambda}{2})^{n-1}\right]\left[(1-\frac{\lambda}{2})(1-r)+c_{22}(r-1)\right]+(1-\frac{\lambda}{2})r\right]}{\left[1-(1-\frac{\lambda}{2})^n\right]\left[1-r+(1-\lambda)(1-\frac{\lambda}{2})^n(r-1)\right]+r},$$

where

$$c_{22} = \frac{\theta(1-\lambda)(1-\frac{\lambda}{2})^n-\frac{\lambda}{2}(1-\theta-\frac{\lambda}{2})(1-\theta)^n}{(\theta-\frac{\lambda}{2})}.$$

Therefore, a test of the null hypothesis H_0:$\Pi^{(n)}(\theta) = 1/2$ or $r = 1$ is not equivalent to a test for linkage.

Test of Linkage between a Marker Locus and a Disease Locus

We have proved that $\Pi^{(n)}(\theta)$ is a strict monotonic function of θ for both the IA and the CGF models when $r \neq 1$. Furthermore, $\Pi^{(n)}(\theta)$ achieves its maximum or minimum at the disease locus. Therefore, a way to test for linkage between a marker locus and a disease locus is to test the null hypothesis $\Pi^{(n)}(\theta) = \Pi^{(n)}(0.5)$, which is equivalent to testing $\theta = 0.5$ when $r \neq 1$. That is, we can test for linkage by testing that the parental ancestry of the marker alleles is equal to that of a marker unlinked to the disease locus. Assume that we can estimate $\Pi^{(n)}$ at any position in the genome. For simplicity, we further assume, for the moment, that there is only one disease locus across the genome. Then, the vast majority of the markers across the genome are unlinked to the disease locus. Theoretically, at all marker locations unlinked to the disease locus, $\Pi^{(n)}$ is expected to equal $\Pi^{(n)}(0.5)$, so we can estimate the distribution of $\Pi^{(n)}$ under the null hypothesis by genotyping markers unlinked to the disease locus and calculating $\Pi^{(n)}$ at these unlinked markers. This procedure is similar to the genomic control method, in which a set of unlinked markers is used to control the false-positive rate due to the effect of population heterogeneity (Devlin and Roeder 1999). For instance, assume that we have estimated proportions of X by descent $\hat{\Pi}_1^{(n)}, \hat{\Pi}_2^{(n)}, \ldots, \hat{\Pi}_m^{(n)}$ at m markers unlinked to the disease locus. Asymptotically, we can assume that $\hat{\Pi}_1^{(n)}, \hat{\Pi}_2^{(n)}, \ldots, \hat{\Pi}_m^{(n)}$ approximate a normal distribution, $N[\Pi^{(n)}(0.5), \sigma^2]$. To test the null hypothesis that a marker is unlinked to the disease locus—that is, $\Pi^{(n)}(\theta) =$

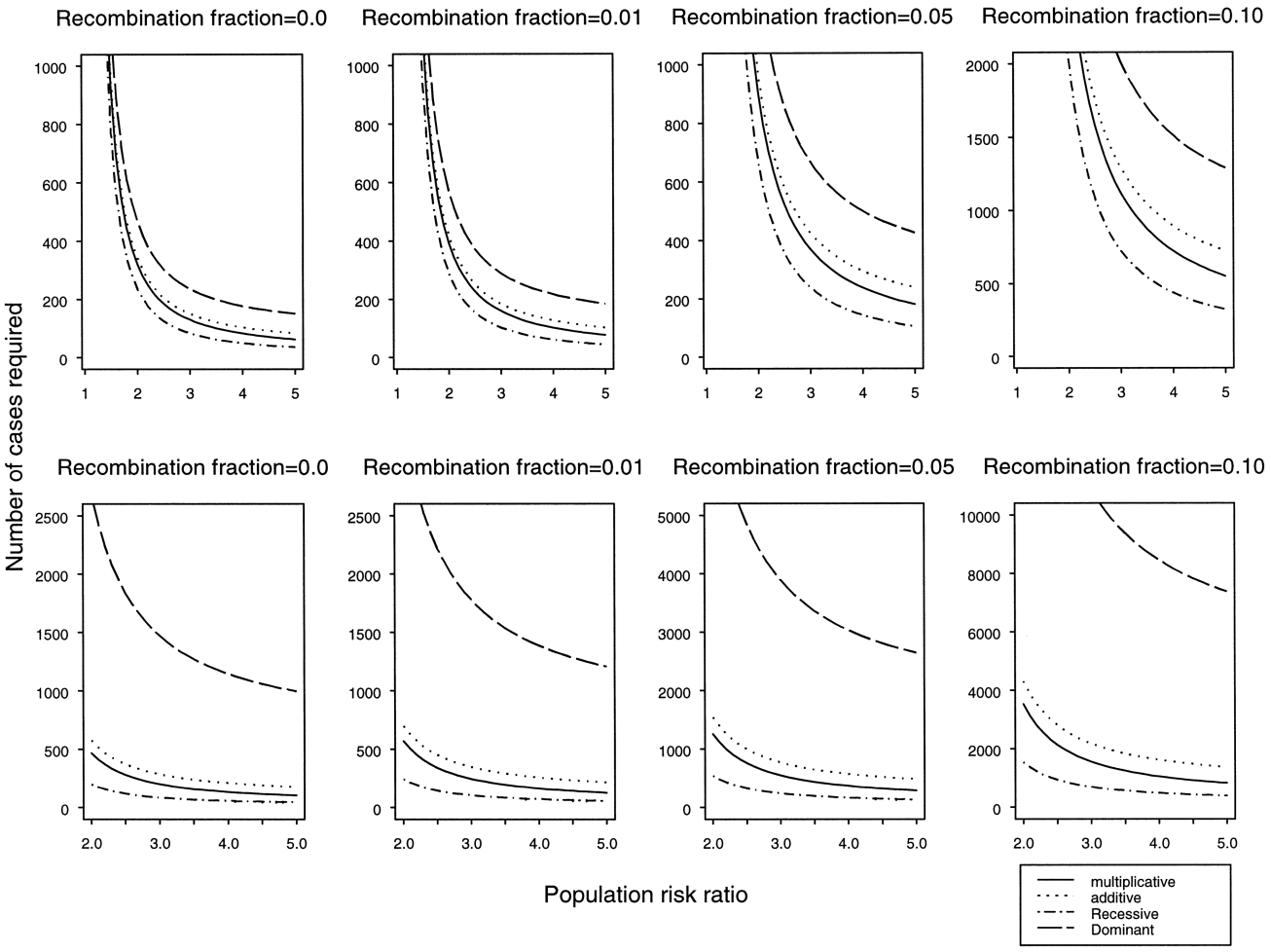

Figure 2 IA model. Number of cases required for 90% power to detect linkage at a significance level of .001 at different population risk ratios and recombination fractions between marker and disease loci under four genetic models: multiplicative, additive, recessive (with the low-risk allele in the low-risk population), and dominant (with the high-risk allele in the high-risk population). *Upper panels,* Total contributions of the parental populations X and Y are 50%/50% ($\lambda = 1.0$). *Lower panels,* Total contributions of parental populations X and Y are 74%/26% ($\lambda = 0.52$). Samples are drawn from the 10th generation.

$\Pi^{(n)}(0.5)$, if $\Pi^{(n)}(0.5)$ and σ^2 are known—we can use the statistic

$$Z = \frac{\hat{\Pi}^{(n)}(\theta) - \Pi^{(n)}(0.5)}{\sigma}$$

and assume that Z follows a standard normal distribution, where $\hat{\Pi}^{(n)}(\theta)$ is the estimator of $\Pi^{(n)}(\theta)$ at the marker locus tested. On the assumption that we know which population has the higher disease risk due to segregation at the locus being tested, so that we know a priori the sign of Z, we perform a one-sided test and reject the null hypothesis at the α significance level if $|Z|$ is greater than the $100(1-\alpha)$th percentile of the standard normal distribution. In practice, we would estimate $\Pi^{(n)}(0.5)$ and σ^2 from $\hat{\Pi}_1^{(n)}, \hat{\Pi}_2^{(n)}, \ldots, \hat{\Pi}_m^{(n)}$ by the sample mean and variance and perform the corresponding one-sided t test.

The proposed test requires estimating $\Pi^{(n)}$ conditional on the marker data. In appendix B, we propose a hidden Markov model (HMM) method to estimate $\Pi^{(n)}$. We refer to the HMM method that uses the transition matrix derived from the IA model as "IA-HMM," and we refer to that derived from the CGF model as "CGF-HMM." As a by-product, we can also estimate the number of generations since the admixture occurred and the current admixture rate. The proposed test also requires markers unlinked to the disease locus, which we cannot observe with certainty. However, we can include all of the available markers, because including markers linked to the disease locus will only lead to the test being conservative—the estimated $\Pi^{(n)}(0.5)$ will be biased toward

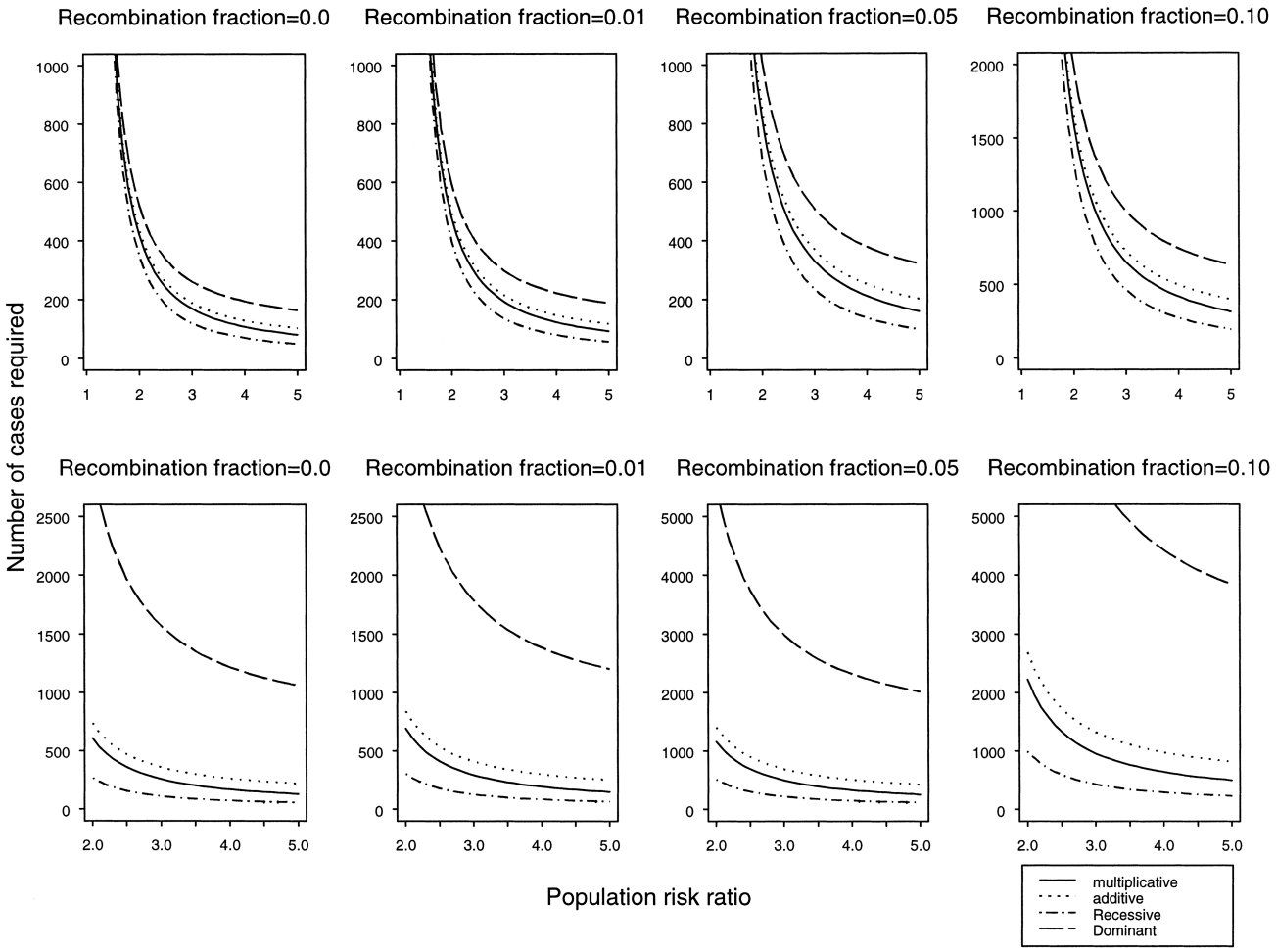

Figure 3 CGF model. Number of cases required for 90% power to detect linkage at a significance level of .001 at different population risk ratios and recombination fractions between marker and disease loci under four genetic models: multiplicative, additive, recessive (with the low-risk allele in the low-risk population), and dominant (with the high-risk allele in the high-risk population). *Upper panels,* Total contributions of parental populations X and Y are 50%/50% ($\lambda = 1$). *Lower panels,* total contributions of parental population X and Y are 74%/26% ($\lambda = 0.06$). Samples are drawn from the 10th generation.

the alternative hypothesis, and the variance will be increased. Alternatively, especially if there may be several linked loci, we can assume a mixture of two distributions and use a commingling analysis (MacLean et al. 1976; Efron 2004) to obtain more-appropriate estimates of $\Pi^{(n)}(0.5)$ and σ^2.

Power of Admixture Mapping

The power to detect linkage between a marker and a disease locus is dependent on various model parameters, including penetrance functions, disease allele frequencies in X and Y, different admixture rates, the recombination fraction between the marker and disease loci, and the accuracy of the estimated proportion X by descent at a marker locus. Assume that we can accurately estimate this ancestral probability at a marker locus. Then, we can model drawing a locus from the ancestral populations X and Y with a binomial distribution (McKeigue 1998). With a one-sided type I error rate α and power of detecting linkage $1 - \beta$, for large samples the required number of cases given the probability $\Pi^{(n)}(\theta)$ is then

$$N \geqslant \frac{1}{2}\left(\frac{\sqrt{\Pi^{(n)}(0.5)[1-\Pi^{(n)}(0.5)]}Z_{1-\alpha} + \sqrt{\Pi^{(n)}(\theta)[1-\Pi^{(n)}(\theta)]}Z_{1-\beta}}{\Pi^{(n)}(\theta) - \Pi^{(n)}(0.5)}\right)^2,$$

where the factor 1/2 arises from each individual having two alleles.

Figure 2 presents the sample size required at the 10th generation for the IA model, under four possible modes

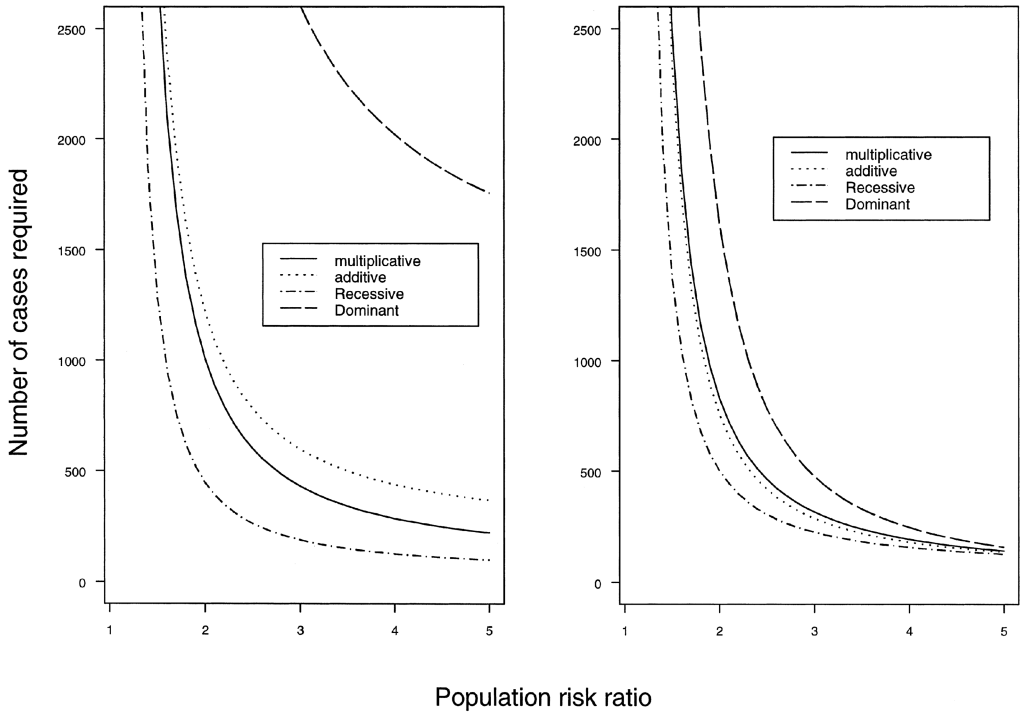

Figure 4 Number of cases required for 90% power to detect linkage at a genomewide significance level of .05, when 3,000 markers are evenly placed along the genome, under four genetic models: multiplicative, additive, recessive (with the low-risk allele in the low-risk population), and dominant (with the high-risk allele in the high-risk population). The population has been admixed for 10 generations, according to the CGF model. Total contributions of parental populations X and Y are 74%/26% ($\lambda = 0.06$). *Left,* X has a higher population risk than Y; the curve for the dominant model does not appear because the number is >2,500. *Right,* Y has a higher population risk than X. Samples are drawn from the 10th generation.

of genetic inheritance: multiplicative, additive, recessive with a low-risk allele in the low-risk population, and dominant with a high-risk allele in the high-risk population, as studied by McKeigue (1998). The sample size is calculated to provide 90% power to detect linkage at a significance level of .001. The sample sizes required at the 10th generation for the CGF model, under the same genetic inheritance models, are presented in figure 3. In both figures 2 and 3, the upper row of panels shows results calculated with a current admixture rate of 50%/50% (X/Y), which corresponds to $\lambda = 1.0$ for the IA model and $\lambda = 0.06$ for the CGF model, respectively. The bottom row of panels shows results calculated with $\lambda = 0.52$ and 0.03 for the IA and CGF models, respectively, corresponding to an admixture rate of 74%/26% (X/Y), which mimics the contemporary African American population. We assign parental population X a higher disease risk than population Y. These results suggest that power is at a maximum for admixture mapping when the admixture rates are equal. The sample size required for the CGF model is slightly larger than that required for the IA model when $\theta < 0.01$ but is smaller when $\theta \geqslant 0.05$, especially for the dominant model. Overall, 1,000 cases are enough to detect linkage with 90% power at a significance level of .001 when the population relative risk $r > 2$ and $\theta < 0.05$, except under a dominant inheritance model. For the CGF model, this sample size is sufficient when the current admixture rate is similar to what would be found when $n = 20$ generations (data not shown). If $\theta = 0$, which indicates that the marker and disease loci are at the same position, 500 cases are usually enough—again, except under dominant inheritance. In contrast, to achieve 90% power at a significance level of .001 to detect linkage, for the sample size of the affected-sib-pair design to be <500 families requires the frequency of the high-risk allele to be <0.25 or the genotype relative risk ratio to be >3 (Risch and Merikangas 1996; McKeigue 1998).

If we consider genomewide admixture mapping with 1 marker/cM and a total of 3,000 markers genotyped, what is the sample size required to achieve 90% power at a significance level of .05? Assume that the disease locus is halfway between two adjacent markers. With Bonferroni correction, a single test must attain the 1.67×10^{-5} significance level. Figure 4 shows the sam-

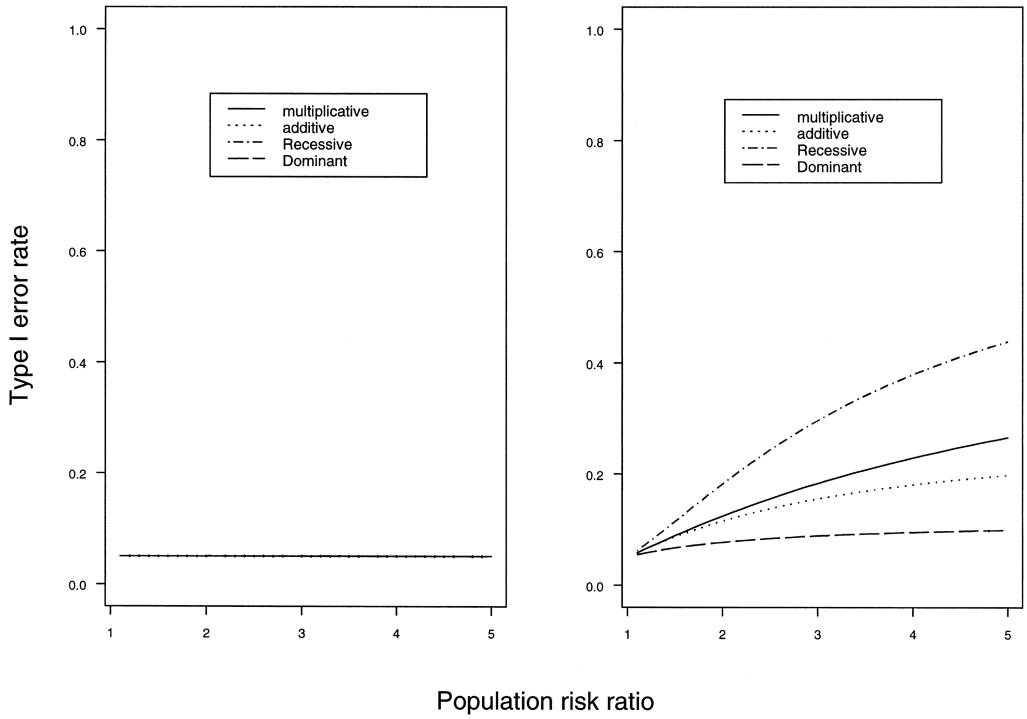

Figure 5 Type I error rate when we test linkage by testing $r = 1$ under four genetic models: multiplicative, additive, recessive (with the low-risk allele in the low-risk population), and dominant (with the high-risk allele in the high-risk population). Total contributions of the parental populations X and Y are 74%/26%. *Left,* IA model. *Right,* CGF model. Two hundred fifty cases are drawn from the 10th generation.

ple size required to reach this level for the CGF model with an admixture rate of 0.74/0.26. The left panel represents the situation in which parental population X has a higher population risk than does parental population Y, and the right panel represents that when Y has a higher population risk than X. In the case of higher risk in Y, usually no more than 1,000 cases are required in order to have 90% power to detect the gene, provided that $r > 2$, except under the dominant model. If X has a higher risk than Y, the sample size needs to be increased in order to achieve the same power. It is apparent that admixture mapping has poorer power under a dominant model. The dominant model requires only one copy of the disease allele to evaluate the risk, resulting in less departure of the ancestral proportion from an unlinked region at a disease marker locus. Thus, to have the same power, a dominant model requires a larger sample size than do other models.

We also calculated the type I error rate when using McKeigue's method of testing $r = 1$ under the true null hypothesis $\theta = 0.5$. Figure 5 presents the type I error for both the IA and CGF models when the current admixture rate is 74%/26% and $n = 10$. The type I error rate is not inflated for the IA model (*left panel*), but it is for the CGF model (*right panel*).

Simulation Studies

To validate the proposed method, we conducted simulation studies. We assumed that there were two parental populations, X and Y. The allele frequencies of Biaka, extracted from ALFRED (Cheung et al. 2000), were used as the marker allele frequencies in X. For simplicity, we converted the microsatellite markers to diallelic markers by pooling the rarer alleles. The marker frequency in Y was taken to be the corresponding allele frequency in X after adding/subtracting 0.4, according to whether the allele frequency in X was less/greater than 0.5, and then adding a random value drawn from a uniform distribution between 0 and 0.1. We assumed that the markers were evenly distributed across the genome. Samples were then simulated according to the admixture models in figure 1. In brief, at the first generation, the marker genotypes of 10,000 unrelated people were simulated according to the marker allele frequencies in population X under the assumption of HWE and independence of the markers. An admixed population was then formed by taking a proportion λ randomly selected from population X to marry with people generated according to the marker allele frequencies in population Y, with the remaining proportion, $1 - \lambda$, randomly mating among

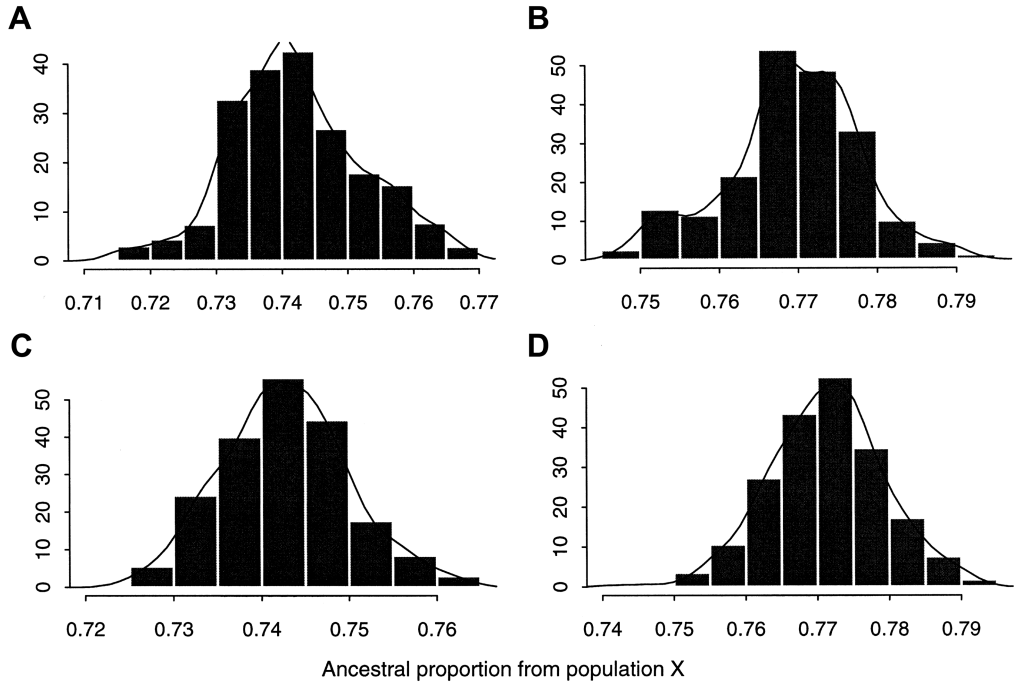

Figure 6 Histograms of estimated ancestral proportions (Π) across the markers; 1,000 markers on 1,000 individuals were generated. *A*, Data simulated using the IA model at a marker density of 1 marker/cM. *B*, Data simulated using the CGF model at a marker density of 1 marker/cM. *C*, Data simulated using the IA model at a marker density of 1 marker/5cM. *D*, Data simulated using the CGF model at a marker density of 1 marker/5 cM.

themselves. The number of children produced by each marriage was assumed to follow a Poisson distribution with mean size 2. The number of crossovers between two marker loci at a distance d (in cM) was assumed to follow a Poisson distribution with mean $d/100$. This process was repeated in the following generations. We let λ equal 0 after the second generation for the IA model and equal a constant value in all generations for the CGF model. All of the samples were drawn from the 10th generation. To simulate which individuals were affected, we let the first marker be the disease locus and calculated the penetrance functions according to the population risk ratio, the allele frequencies in the two parental populations X and Y, and the inheritance model. In the following simulations, we use IA-HMM to estimate the ancestral proportions for data generated under the IA model and use CGF-HMM for the data generated under the CGF model.

Is $\hat{\Pi}_1^{(n)},\hat{\Pi}_2^{(n)},\ldots,\hat{\Pi}_m^{(n)}$ *Asymptotically Normally Distributed?*

Since the estimates of $\hat{\Pi}_1^{(n)},\hat{\Pi}_2^{(n)},\ldots,\hat{\Pi}_m^{(n)}$ are dependent on the marker informativeness for ancestry, the distribution of $\hat{\Pi}_1^{(n)},\hat{\Pi}_2^{(n)},\ldots,\hat{\Pi}_m^{(n)}$ may depart from a normal distribution. We therefore randomly sampled 1,000 unrelated individuals from the 10th generation, for both the IA and CGF models, and estimated $\hat{\Pi}_1^{(n)},\hat{\Pi}_2^{(n)},\ldots,\hat{\Pi}_m^{(n)}$, using the proposed HMM method separately for the IA and CGF models. Figure 6 presents histograms of the estimates $\hat{\Pi}_1^{(n)},\hat{\Pi}_2^{(n)},\ldots,\hat{\Pi}_m^{(n)}$, where 1,000 markers were evenly spaced at a density of 1 marker/cM and a density of 1 marker/5 cM. The results suggest that there can be departure from the normal density function. The degree of the departure is dependent on the marker density, with the greater marker density resulting in more departure from the normal distribution. The reason for this is that the closer two markers are, the stronger their correlation. However, the correlation can be reduced by selecting the estimates of Π at marker loci every 5 cM.

We next compared a person's true ancestral proportion from population X to its estimate when the HMM is used. The upper panels in figure 7 present scatterplots of the true and estimated proportions for different marker densities and admixture models when ancestral marker allele frequencies are known. The results suggest that the HMM can accurately estimate the ancestral proportion when markers are spaced <5 cM apart, provided that the ancestral marker allele frequencies are known. To further explore the effect of marker density on the estimates of ancestral proportions, we calculated the correlation and the average difference between the true and the estimated ancestral proportions at the density of one marker every 1 cM, 2 cM, and 5 cM (table 2). The

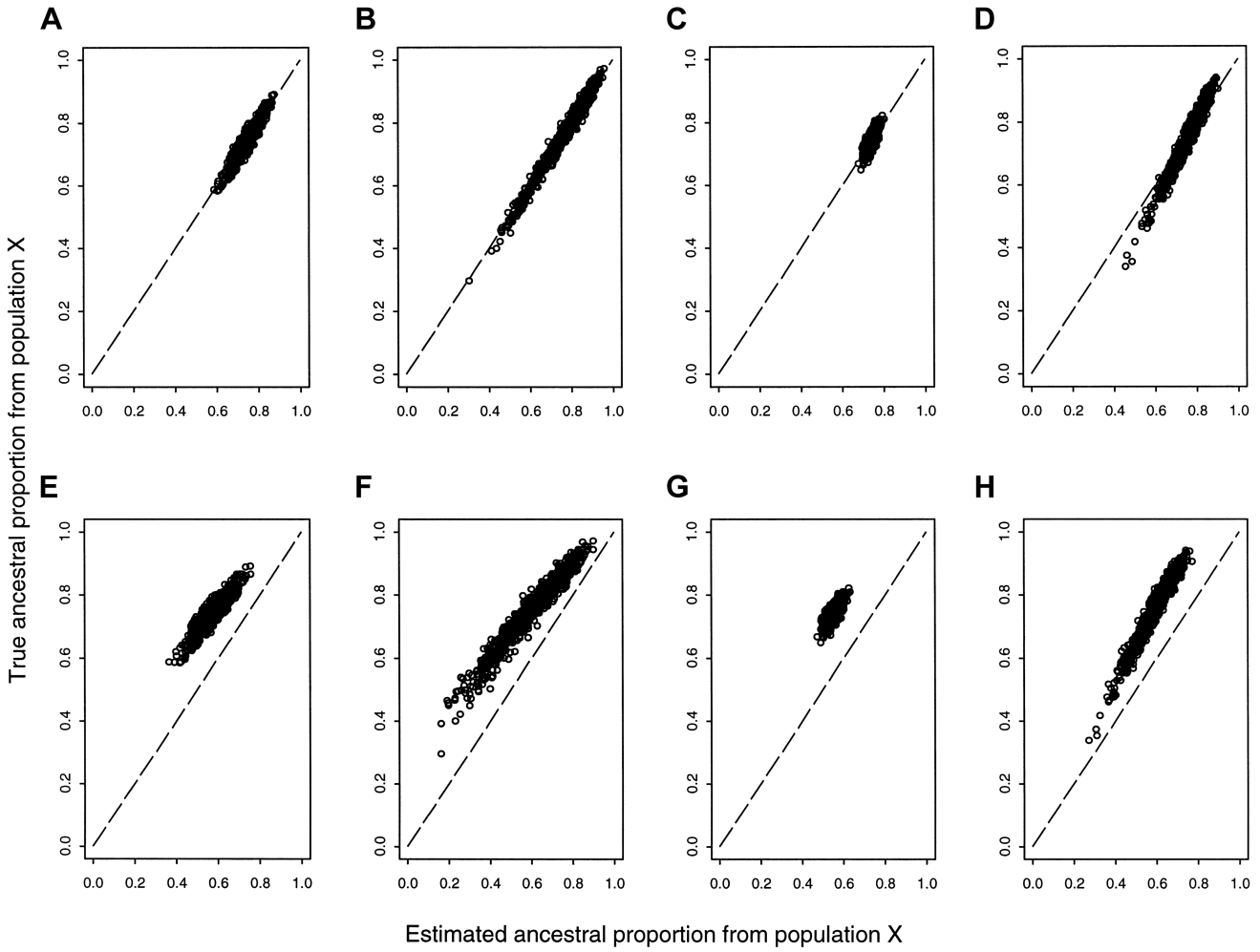

Figure 7 Comparisons of a person's estimated and true ancestral proportion; 1,000 markers were generated on 1,000 individuals. *A–D*, Known ancestral allele frequencies. *E–H*, Ancestral allele frequencies estimated by STRUCTURE. *A* and *E*, Data simulated using the IA model at a marker density of 1 marker/cM. *B* and *F*, Data simulated using the CGF model at a marker density of 1 marker/cM. *C* and *G*, Data simulated using the IA model at a marker density of 1 marker/5 cM. *D* and *H*, Data simulated using the CGF model at a marker density of 1 marker/5 cM.

correlation decreases and the average difference increases as the marker spacing increases. We can also observe that the HHM gives an unbiased estimate of the ancestral proportion. Our results suggest that the performance of the HMM at a density of one marker every 2 cM is almost as good as that at a density of one marker every 1 cM. Furthermore, when we applied the CGF-HMM to data simulated on the basis of the IA model, we obtained essentially the same results as that from applying the IA-HMM, indicating that the CGF-HMM can perform well for data generated from either the IA or the CGF model.

In practice, we seldom know the true ancestral allele frequencies. To explore how the ancestral allele frequencies affect the performance of the HMM, we first estimated these allele frequencies through use of the unlinked model in STRUCTURE (Falush et al. 2003). These estimated allele frequencies were then used in the HMM. The bottom panels of figure 7 present the results from the HMM for the same data as the upper panels. Table 2 also presents the correlation and the average difference when ancestral allele frequencies are unknown. We observed that the true ancestral proportions were systematically underestimated.

Power Simulations

We next performed simulations to compare the theoretical power with the empirical power. Because the theoretical sample size required to reach genomewide significance is much larger for a dominant mode of inheritance, our simulations focused on the multiplicative,

Table 2

The Correlation and Difference (Average ± SD) between the True and Estimated Ancestral Proportions

Model and Marker Distance (in cM)	Ancestral Frequencies Known		Ancestral Frequencies Unknown	
	Correlation	Difference	Correlation	Difference
IA:				
1	.947	$1.9 \times 10^{-4} \pm .016$	.926	$-.168 \pm .023$
2	.892	$-1.6 \times 10^{-3} \pm .017$	.873	$-.183 \pm .019$
5	.776	$2.4 \times 10^{-4} \pm .016$	.763	$-.189 \pm .017$
CGF:				
1	.99	$8.6 \times 10^{-5} \pm .015$	.977	$-.145 \pm .039$
2	.986	$8.7 \times 10^{-4} \pm .018$	.979	$-.16 \pm .023$
5	.982	$1.8 \times 10^{-3} \pm .028$	.976	$-.166 \pm .025$

additive, and recessive models. For each mode of inheritance and a given population relative risk, we simulated the number of cases predicted, in theory, to have 90% power for 5% genomewide significance based on the CGF model. We then estimated the ancestral proportions using the CGF-HMM and calculated the empirical power, using a one-sided Z test. Table 3 presents the power when the marker density was 1 marker/cM, based on 100 replications. The empirical power was slightly higher than the theoretical power. When the marker density was 1 marker/2 cM, the results were similar (data not shown). We also calculated the type I error based on markers unlinked to the disease locus, and that was also within the nominal level (table 3). The average number of generations was estimated to be 9.9 ± 0.31 for the additive model when the population relative risk was 2 and was similar for the other modes of inheritance.

Discussion

Pfaff et al. (2001) explored, via computer simulation, the distribution of LD created by recent population admixture for both the IA and CGF models. They concluded that admixture mapping can be confounded by complex population and evolutionary history, resulting in a high false-positive rate. Our theoretical work further demonstrates that the method conditional on parental admixture (McKeigue 1998) is not testing for linkage, as was originally claimed. With this method, the presence of linkage is tested only if the population is equally admixed and the admixture occurs in a single generation. For the IA model, the type I error of this method is reasonable as long as the admixture occurred ~10 generations ago. However, the type I error will be inflated when the admixture takes place within 5 generations (data not shown). For the CGF model, even if admixture occurred 10 generations ago, the type I error can still be inflated. The reason for this is that gametic association created by the population admixture process alone cannot disappear within 5 generations for the IA model and will last forever for the CGF model because of the admixture at every generation.

The admixture mapping method can be ameliorated by considering the distribution of ancestral probabilities at unlinked marker loci as the null distribution under the hypothesis of no linkage. This approach makes it possible to test for linkage with controlled type I error. It should be noted that we need only sample cases in the admixed population, an advantage over the case-control design. Except for under the dominant model, the power to detect linkage with this method is generally adequate, since we have shown that the sample size to detect genomewide 5% significant linkage with 90% power is within a practical range, provided the population relative risk ratio is >2. In some circumstances, this method is more powerful than affected-sib-pair linkage analysis. In both the IA and CGF models, we consider recombination as the only source of disruption of the LD between the two loci. In practice, other factors, such as mutation and natural selection, may affect the power of the proposed admixture mapping method. However, the effect of these factors may be expected to be limited because of the short interval of elapsed time.

A challenge of the proposed approach is to select the markers and estimate the ancestral probability of the marker loci among cases. Although the markers for a current genome-scan linkage analysis, usually with a density of 1 marker/10 cM, may be applicable for admixture mapping, they may not provide enough information to estimate the ancestral probability accurately. In practice, markers need to be selected according to the marker information content for ancestry, as measured by Wright's F_{ST}, or by a more reasonable measure recently proposed by Rosenberg et al. (2003). In view of the results of our simulations, we suggest using 1,500–3,000 informative markers in a genomewide admixture mapping study. It has been pointed out by McKeigue (1998) that, with multipoint statistical methods, we can extract 80% of the information regarding ancestry with a marker spacing of 1 cM in populations

Table 3

Empirical Power and Type I Error Rates, for Different Modes of Inheritance and Population Risk Ratios, for Theoretically Calculated Sample Sizes to Give 90% Power at the 5% Significance Level

Mode of Inheritance	Power for Genomewide Significance[a] at a Population Risk Ratio of			Type I Error Rate[b] at a Population Risk Ratio of		
	2	3	5	2	3	5
Multiplicative	.92	.93	.97	.0501	.0505	.0478
Additive	.94	.92	.83	.0472	.0463	.0486
Recessive	.97	.96	.93	.0506	.0489	.0454

NOTE.—Cases were drawn from the 10th generation according to the CGF model. Ancestral proportions were also estimated using a HMM based on the CGF model.

[a] Power is based on 100 replications.

[b] Type I error rate was calculated as the total number of markers unlinked to the disease locus with test statistic $|Z| > 1.645$ (one-sided test) divided by the total number of such markers.

in which admixture has occurred <10 generations ago. It may be difficult to reach such a density by selecting informative markers from among microsatellites alone (Collins-Schramm et al. 2002; Smith et al. 2001). With their abundant identification across the human genome, we can be more flexible in selecting SNPs for admixture mapping (Sachidanandam et al. 2001). Furthermore, it has been demonstrated that much of the genome can be parsed into long "blocks," within which little recombination has occurred (Gabriel et al. 2002). These blocks are separated by small regions in which recombination "hotspots" are located, suggesting an efficient approach to admixture mapping. Zhu et al. (2003) found that African Americans have more haplotype blocks than European Americans in the RAS gene, and the haplotype blocks of African Americans are usually subintervals of those among European Americans. This result is also consistent with the large survey by Gabriel et al. (2002). Wright's F_{ST} can be increased if haplotypes in a block are considered as marker alleles (Zhu et al. 2003). Thus, using SNPs can be a feasible approach to marker selection for admixture mapping. We can also view haplotype blocks as unbroken units transmitted from one or the other ancestral population. However, the role of hotspots in forming haplotype blocks is still in debate (Wall and Pritchard 2003), and further research should be done to explore how useful haplotype blocks are for admixture mapping. To estimate the admixture of an individual at a particular locus, we have extended (appendix B) the HMM method proposed by McKeigue (1998), who used a two-state Markov process. In our method, we use a three-state Markov process that can be directly applied to a set of genotypes instead of to individual chromosomes. Therefore, our HMM method can extract more information by allowing for uncertainty, because we do not need to reconstruct haplotypes. Our simulations suggest that this method works well in general.

Finally, our proposed method requires knowing the allele frequencies in the parental populations. Although we may obtain most of the allele frequencies in parental populations from current databases, these frequencies may only approximate those in the true parental populations of our studied sample. For example, we may have difficulty finding the true allele frequencies for the appropriate African population when we study African Americans. This may create a potential limitation to admixture mapping. As an alternative, we may first apply STRUCTURE (Falush et al. 2003) to estimate the ancestral allele frequencies and then input these estimated frequencies into our model. Our simulations suggest that the estimate of ancestral proportions may be biased in the same direction for each position, resulting in much of the bias being eliminated when we compare the ancestral proportion at one locus with that in a region. Further research should be done to explore the effect of this on type I error.

Acknowledgments

We thank Dr. S. L. Zhang, for providing a C program to generate admixed samples, and three reviewers, for their constructive comments on the initial version of this article. This work was supported by National Heart, Lung and Blood Institute grants HL54466 and HL65702, National Center for Research Resources grant RR03655, National Institute of Diabetes, Digestive and Kidney Diseases grant DK57292, and National Institute of General Medicine Sciences grant GM28356.

Appendix A

We use the same notation as in the text. According to the CGF model assumption, the haplotype frequencies produced in generation 0 are $h_1^{(0)} = p_X$, $h_2^{(0)} = q_X$, $h_3^{(0)} = 0$, and $h_4^{(0)} = 0$. Let $g_1^{(n)}, g_2^{(n)}, \ldots, g_{10}^{(n)}$ represent the frequencies of the 10 genotypes at generation n (table 1). These genotypes arise from a $1 - \lambda$ part coming from "self mating" in parental population X and a λ part coming from admixture between parental populations X and Y. Thus, we can write the 10 genotype probabilities as in table A1.

Let $\Delta_1^{(n)} = h_1^{(n)}h_4^{(n)} - h_2^{(n)}h_3^{(n)}$, $\Delta_2^{(n)} = h_1^{(n)}q_Y - h_2^{(n)}p_Y$, and $H^{(n)T} = (h_1^{(n)}, h_2^{(n)}, h_3^{(n)}, h_4^{(n)}, \Delta_1^{(n)}, \Delta_2^{(n)})$, where the superscript "T" represents "transpose." Then, $H^{(0)T} = (p_X, q_X, 0, 0, 0, p_X - p_Y)$. By using table 1, letting subscripts denote the dimen-

Table A1

Genotype Probabilities for the IA and CGF Models in Generation *n*

	Genotype Probability under Model									
Model	$\frac{M_X}{D}\|\frac{M_X}{D}$ (g_1)	$\frac{M_X}{D}\|\frac{M_X}{d}$ (g_2)	$\frac{M_X}{d}\|\frac{M_X}{d}$ (g_3)	$\frac{M_X}{D}\|\frac{M_Y}{D}$ (g_4)	$\frac{M_X}{D}\|\frac{M_Y}{d}$ (g_5)	$\frac{M_X}{d}\|\frac{M_Y}{D}$ (g_6)	$\frac{M_X}{d}\|\frac{M_Y}{d}$ (g_7)	$\frac{M_Y}{D}\|\frac{M_Y}{D}$ (g_8)	$\frac{M_Y}{D}\|\frac{M_Y}{d}$ (g_9)	$\frac{M_Y}{d}\|\frac{M_Y}{d}$ (g_{10})
IA	$h_1^{(n-1)}h_1^{(n-1)}$	$2h_1^{(n-1)}h_2^{(n-1)}$	$h_2^{(n-1)}h_2^{(n-1)}$	$2h_1^{(n-1)}h_3^{(n-1)}$	$2h_1^{(n-1)}h_4^{(n-1)}$	$2h_2^{(n-1)}h_3^{(n-1)}$	$2h_2^{(n-1)}h_4^{(n-1)}$	$h_3^{(n-1)}h_3^{(n-1)}$	$2h_3^{(n-1)}h_4^{(n-1)}$	$h_4^{(n-1)}h_4^{(n-1)}$
CGF	$(1-\lambda)$ $\times h_1^{(n-1)}h_1^{(n-1)}$	$2(1-\lambda)$ $\times h_1^{(n-1)}h_2^{(n-1)}$	$(1-\lambda)$ $\times h_2^{(n-1)}h_2^{(n-1)}$	$2(1-\lambda)$ $\times h_1^{(n-1)}h_3^{(n-1)}$ $+\lambda h_1^{(n-1)}p_Y$	$2(1-\lambda)$ $\times h_1^{(n-1)}h_4^{(n-1)}$ $+\lambda h_1^{(n-1)}q_Y$	$2(1-\lambda)$ $\times h_2^{(n-1)}h_3^{(n-1)}$ $+\lambda h_2^{(n-1)}p_Y$	$2(1-\lambda)$ $\times h_2^{(n-1)}h_4^{(n-1)}$ $+\lambda h_2^{(n-1)}q_Y$	$(1-\lambda)$ $\times h_3^{(n-1)}h_3^{(n-1)}$ $+\lambda h_3^{(n-1)}p_Y$	$2(1-\lambda)$ $\times h_3^{(n-1)}h_4^{(n-1)}$ $+\lambda h_3^{(n-1)}q_Y$ $+\lambda h_4^{(n-1)}p_Y$	$(1-\lambda)$ $\times h_4^{(n-1)}h_4^{(n-1)}$ $+\lambda h_4^{(n-1)}q_Y$

sions of a matrix and letting I be the identity matrix, we can obtain the following iterative equation: $H^{(n)} = AH^{(n-1)} + D$, where

$$A = \begin{bmatrix} \left(1-\frac{\lambda}{2}\right)I_{4\times 4} & B_{4\times 2} \\ 0_{2\times 4} & C_{2\times 2} \end{bmatrix}$$

and

$$D^T = \begin{bmatrix} 0 & 0 & \frac{\lambda p_Y}{2} & \frac{\lambda q_Y}{2} & 0 & 0 \end{bmatrix},$$

where $B_{4\times 2}$ and $C_{2\times 2}$ are

$$B_{4\times 2} = \begin{bmatrix} -1+\lambda & -\frac{\lambda}{2} \\ 1-\lambda & \frac{\lambda}{2} \\ 1-\lambda & \frac{\lambda}{2} \\ -1+\lambda & -\frac{\lambda}{2} \end{bmatrix}\theta$$

and

$$C_{2\times 2} = \begin{bmatrix} \left(1-\frac{\lambda}{2}\right)^2-(1-\lambda)\theta & \frac{\lambda}{2}\left(1-\frac{\lambda}{2}-\theta\right) \\ -(1-\lambda)\theta & 1-\frac{\lambda}{2}-\frac{\lambda\theta}{2} \end{bmatrix}.$$

Ignoring the subscripts will not result in confusion; therefore, we can write $H^{(n)} = A^nH^{(0)} + (I-A)^{-1}(I-A^n)D$, where

$$A^n = \begin{bmatrix} \left(1-\frac{\lambda}{2}\right)^n I & \left(1-\frac{\lambda}{2}\right)^{n-1}B\left[I-\left(1-\frac{\lambda}{2}\right)^{-1}C\right]^{-1}\left[I-\left(1-\frac{\lambda}{2}\right)^{-n}C^n\right] \\ 0 & C^n \end{bmatrix}.$$

Rewriting matrix C as $C = C_1C_2$, where

$$C_1 = \begin{bmatrix} \left(1 - \frac{\lambda}{2}\right)^2 & 1 \\ 0 & 1 \end{bmatrix}$$

and

$$C_2 = \begin{bmatrix} 1 & -1 \\ -(1-\lambda)\theta & 1 - \frac{\lambda}{2} - \frac{\lambda\theta}{2} \end{bmatrix} ,$$

implies $C^n = C_1(C_2C_1)^{n-1}C_2$. Some algebra then leads to

$$C^n = \left(1 - \frac{\lambda}{2}\right)^{n-1} \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} ,$$

where

$$c_{11} = \frac{(1-\lambda)\theta(1-\theta)^n - \frac{\lambda}{2}(1-\theta-\frac{\lambda}{2})(1-\frac{\lambda}{2})^n}{\theta - \frac{\lambda}{2}} ,$$

$$c_{12} = \frac{\frac{\lambda}{2}(1-\theta-\frac{\lambda}{2})[(1-\frac{\lambda}{2})^n - (1-\theta)^n]}{\theta - \frac{\lambda}{2}} ,$$

$$c_{21} = \frac{-(1-\lambda)\theta[(1-\frac{\lambda}{2})^n - (1-\theta)^n]}{\theta - \frac{\lambda}{2}} ,$$

and

$$c_{22} = \frac{\theta(1-\lambda)(1-\frac{\lambda}{2})^n - \frac{\lambda}{2}(1-\theta-\frac{\lambda}{2})(1-\theta)^n}{\theta - \frac{\lambda}{2}} .$$

Therefore, we obtain

$$h_1^{(n)} = \left(1 - \frac{\lambda}{2}\right)^n p_Y + c_{22}(p_X - p_Y) ,$$

$$h_2^{(n)} = \left(1 - \frac{\lambda}{2}\right)^n q_Y + c_{22}(q_X - q_Y) ,$$

$$h_3^{(n)} = \left[\left(1 - \frac{\lambda}{2}\right)^n - c_{22}\right](p_X - p_Y) + \left[1 - \left(1 - \frac{\lambda}{2}\right)^n\right]p_Y ,$$

and

$$h_4^{(n)} = \left[\left(1-\frac{\lambda}{2}\right)^n - c_{22}\right](q_X - q_Y) + \left[1-\left(1-\frac{\lambda}{2}\right)^n\right]q_Y \ .$$

Then,

$$\begin{aligned} P(\text{affected}) = {} & f_2[(1-\lambda)(h_1^{(n-1)} + h_3^{(n-1)})^2 + \lambda(h_1^{(n-1)} + h_3^{(n-1)})p_Y] \\ & + f_1[2(1-\lambda)(h_1^{(n-1)} + h_3^{(n-1)})(h_2^{(n-1)} + h_4^{(n-1)}) + \lambda(h_1^{(n-1)} + h_3^{(n-1)})q_Y + \lambda(h_2^{(n-1)} + h_4^{(n-1)})p_Y] \\ & + f_0[(1-\lambda)(h_2^{(n-1)} + h_4^{(n-1)})^2 + \lambda(h_2^{(n-1)} + h_4^{(n-1)})q_Y] \end{aligned}$$

and

$$\begin{aligned} \Pi^{(n)}(\theta) = \frac{1}{P(\text{affected})}\Bigg(& f_2 h_1^{(n-1)}\left[(1-\lambda)(h_1^{(n-1)} + h_3^{(n-1)}) + \frac{1}{2}\lambda p_Y\right] + f_1\Big\{h_2^{(n-1)}\left[(1-\lambda)(h_1^{(n-1)} + h_3^{(n-1)}) + \frac{1}{2}\lambda p_Y\right] \\ & + \left[(1-\lambda)(h_2^{(n-1)} + h_4^{(n-1)}) + \frac{1}{2}\lambda q_Y\right]h_1^{(n-1)}\Big\} + f_0 h_2^{(n-1)}\left[(1-\lambda)(h_2^{(n-1)} + h_4^{(n-1)}) + \frac{1}{2}\lambda q_Y\right]\Bigg) \ . \end{aligned}$$

Thus, $\Pi^{(n)}(\theta)$ can be expressed as a function of p_X, p_Y, θ, n, and the penetrance functions. Furthermore, $\Pi^{(n)}(\theta)$ is a function of θ only through c_{22}. To prove that $\Pi^{(n)}(\theta)$ is a strict monotonic function of θ when $r \neq 1$, we need only demonstrate that c_{22} is a strict monotonic function of θ. Let $z = \theta - (\lambda/2)$, so that

$$-\frac{\lambda}{2} \leqslant z \leqslant \frac{1-\lambda}{2} \ .$$

Taking the derivative of c_{22} with respect to θ is the same as taking that of c_{22} with respect to Z. Therefore,

$$\frac{\partial c_{22}}{\partial \theta} = -\frac{\lambda}{2z^2}\left[(1-\lambda)\left(1-\frac{\lambda}{2}\right)^n + (1-\lambda)\left(1-\frac{\lambda}{2}-z\right)^n - nz(1-\lambda-z)\left(1-\frac{\lambda}{2}-z\right)^{n-1}\right] .$$

(1) When $-\lambda/2 \leqslant z < 0$,

$$\frac{\partial c_{22}}{\partial \theta} \leqslant -\frac{\lambda}{2z^2}(1-\lambda)\left(1-\frac{\lambda}{2}\right)^n < 0 \ .$$

(2) When $0 < z \leqslant (1-\lambda)/2$,

$$\frac{\partial c_{22}}{\partial \theta} < -\frac{\lambda(1-\lambda)}{2z^2}\left[\left(1-\frac{\lambda}{2}\right)^n + \left(1-\frac{\lambda}{2}-z\right)^n - nz\left(1-\frac{\lambda}{2}-z\right)^{n-1}\right] .$$

Let $w = z/(1-0.5\lambda)$; then, $0 < w \leqslant 1/2$. Thus,

$$\frac{\partial c_{22}}{\partial \theta} < -\frac{\lambda(1-\lambda)}{2z^2}\left(1-\frac{\lambda}{2}\right)^n\left\{1 + (1-w)^{n-1}[1-(n+1)w]\right\} . \tag{A1}$$

The right side of equation (A1) achieves its maximum when $w = 2/(n+1)$, and so

$$\frac{\partial c_{22}}{\partial \theta} < -\frac{\lambda(1-\lambda)}{2z^2}\left(1-\frac{\lambda}{2}\right)^n\left[1-\left(1-\frac{2}{n+1}\right)^{n-1}\right] < 0 \ ,$$

for all $n > 1$.

When $n = 1$, equation (A1) implies $\partial c_{22}/\partial\theta < 0$. Thus, we have proved that $\Pi^{(n)}$ is a strict monotonic function of θ.

Appendix B

Consider an individual with M ordered marker loci, with known recombination fractions θ_i between loci i and $i + 1$. Under the IA model, we wish to calculate the likelihood for an individual with observed marker genotypes. We apply a method similar to that of Lander and Green (1987) to calculate this likelihood. At a locus M_i, we can have 0, 1, or 2 alleles X by descent. Let v_i be the number of alleles X by descent at the locus M_i. Then $v_1, v_2, \ldots, v_M$ arise from a Markov chain. The transition probabilities depend only on the admixture rate, the number of generations since admixture occurred, and the recombination fractions. To calculate the transition matrix, we consider marker loci M_i and M_{i+1}, separated by recombination fraction θ_i. Let M_{iX} and M_{iY} represent alleles at marker locus M_i X by descent and Y by descent, respectively. At generation 1, we observe the two genotypes $M_{iX}M_{(i+1)X}/M_{iX}M_{(i+1)X}$ and $M_{iX}M_{(i+1)X}/M_{iY}M_{(i+1)Y}$ with frequencies $1 - \lambda$ and λ. Because the mating is random in the following generations, the haplotype frequencies can be calculated without difficulty. The haplotype frequencies generated by generation $n - 1$ are

$$z_1 = p(M_{iX}M_{(i+1)X}) = \left(1 - \frac{\lambda}{2}\right)^2 + \frac{\lambda}{2}\left(1 - \theta - \frac{\lambda}{2}\right)(1 - \theta)^{n-2},$$

$$z_2 = p(M_{iX}M_{(i+1)Y}) = \frac{\lambda}{2}\left[1 - \frac{\lambda}{2} - \left(1 - \theta - \frac{\lambda}{2}\right)(1 - \theta)^{n-2}\right],$$

$$z_3 = p(M_{iY}M_{(i+1)X}) = \frac{\lambda}{2}\left[1 - \frac{\lambda}{2} - \left(1 - \theta - \frac{\lambda}{2}\right)(1 - \theta)^{n-2}\right],$$

and

$$z_4 = p(M_{iY}M_{(i+1)Y}) = \frac{\lambda}{2}\left[\frac{\lambda}{2} + \left(1 - \theta - \frac{\lambda}{2}\right)(1 - \theta)^{n-2}\right].$$

The genotype frequencies at generation n can be easily calculated. Let $T(\theta_i)$ be the transition matrix, so that

$$T(\theta_i) = \begin{bmatrix} \frac{z_4^2}{(z_3 + z_4)^2} & \frac{2z_3z_4}{(z_3 + z_4)^2} & \frac{z_3^2}{(z_3 + z_4)^2} \\ \frac{z_2z_4}{(z_1 + z_2)(z_3 + z_4)} & \frac{z_1z_4 + z_2z_3}{(z_1 + z_2)(z_3 + z_4)} & \frac{z_1z_3}{(z_1 + z_2)(z_3 + z_4)} \\ \frac{z_2^2}{(z_1 + z_2)^2} & \frac{2z_1z_2}{(z_1 + z_2)^2} & \frac{z_1^2}{(z_1 + z_2)^2} \end{bmatrix}.$$

Define a 3×3 diagonal matrix Q_i having rows and columns indexed by l, with elements $q_{ll} = P$(observed genotype at $M_i \mid v_i = l - 1$) and $q_{lj} = 0$ if $l \neq j$. q_{ll} can be easily calculated and is a function of allele frequencies in the parental populations. The likelihood for an individual is then given by

$$L = \delta^T Q_1 T(\theta_1) Q_2 T(\theta_2) \ldots T(\theta_{M-1}) Q_M 1,$$

where δ is a 3×1 vector with elements equal to the prior probabilities of v, which we choose to be $\delta^T = [\lambda^2/4, \lambda(1 - \lambda/2), (1 - \lambda/2)^2]$, and 1 is a 3×1 vector of unities. This likelihood can be used to estimate λ and n. To calculate the expected ancestry at a marker locus conditional on the marker data, we can apply the Lander-Green

algorithm (Lander and Green 1987). Missing genotypes at a marker locus can be easily incorporated into the HMM by directly exploring the next marker locus.

For the CGF model we can use the same method, except for the formula of the transition matrix, which can be calculated similarly.

References

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

Cheung KH, Osier MV, Kidd JR, Pakstis AJ, Miller PL, Kidd KK (2000) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms. Nucleic Acids Res 29:361–363

Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, Hanson RL, Knowler WC, Cooper R, Li H, Seldin MF (2002) Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. Am J Hum Genet 70: 737–750

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J Am Stat Assoc 99:96–104

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296: 2225–2229

Halder I, Shriver MD (2003) Measuring and using admixture to study the genetics of complex disease. Hum Genomics 1: 52–62

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

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

Lautenberger JA, Stephens JC, O'Brien SJ, Smith MW (2000) Significant admixture linkage disequilibrium across 30 cM around the FY locus in African Americans. Am J Hum Genet 66:969–978

Long JC (1991) The genetic structure of admixed populations. Genetics 127:417–428

MacLean CJ, Morton NE, Elston RC, Yee S (1976) Skewness in commingled distributions. Biometrics 32:695–699

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

——— (1998) Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. Am J Hum Genet 63:241–251

McKeigue PM, Carpenter JR, Parra EJ, Shriver MD (2000) Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. Ann Hum Genet 64:171–186

Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E, Shriver MD (2001) Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 68:198–207.

Rife DC (1954) Populations of hybrid origin as source material for the detection of linkage. Am J Hum Genet 6:26–33

Risch N (1992) Mapping genes for complex disease using association studies with recently admixed populations. Am J Hum Genet Suppl 51:13

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

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402–1422

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Smith MW, Lautenberger JA, Shin HD, Chretien JP, Shrestha S, Gilbert DA, O'Brien SJ (2001) Markers for mapping by admixture linkage disequilibrium in African American and Hispanic populations. Am J Hum Genet 69:1080–1094

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 52:506–516

Stephens JC, Briscoe D, O'Brien SJ (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am J Hum Genet 55:809–824

Thomson G (1995) Mapping disease genes: family-based association studies. Am J Hum Genet 57:487–498

Wall JD, Pritchard JK (2003) Assessing the performance of the haplotype block model of linkage disequilibrium. Am J Hum Genet 73:502–515

Zheng C, Elston RC (1999) Multipoint linkage disequilibrium mapping with particular reference to the African-American population. Genet Epidemiol 17:79–101

Zhu X, Yan D, Cooper RS, Luke A, Ikeda MA, Chang YP, Weder A, Chakravarti A (2003) Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program. Genome Res 13:173–181