Am. J. Hum. Genet. 55:809–824, 1994

Mapping by Admixture Linkage Disequilibrium in Human Populations: Limits and Guidelines

J. Claiborne Stephens, David Briscoe,* and Stephen J. O'Brien

Laboratory of Viral Carcinogenesis, National Cancer Institute, National Institutes of Health, Frederick Cancer Research and Development Center, Frederick, MD

Summary

Certain human hereditary conditions, notably those with low penetrance and those which require an environmental event such as infectious disease exposure, are difficult to localize in pedigree analysis, because of uncertainty in the phenotype of an affected patient's relatives. An approach to locating these genes in human cohort studies would be to use association analysis, which depends on linkage disequilibrium of flanking polymorphic DNA markers. In theory, a high degree of linkage disequilibrium between genes separated by 10–20 cM will be generated and persist in populations that have a history of recent (3–20 generations ago) admixture between genetically differentiated racial groups, such as has occurred in African Americans and Hispanic populations. We have conducted analytic and computer simulations to quantify the effect of genetic, genomic, and population parameters that affect the amount and ascertainment of linkage disequilibrium in populations with a history of genetic admixture. Our goal is to thoroughly explore the ranges of all relevant parameters or factors (e.g., sample size and degree of genetic differentiation between populations) that may be involved in gene localization studies, in hopes of prescribing guidelines for an efficient mapping strategy. The results provide reasonable limits on sample size (200–300 patients), marker number (200–300 in 20-cM intervals), and allele differentiation (loci with allele frequency difference of ≥.3 between admixed parent populations) to produce an efficient approach (>95% ascertainment) for locating genes not easily tracked in human pedigrees.

Introduction

The combination of DNA polymorphisms, pedigree analyses of clinical expression, and positional cloning has led to recent discoveries of several human genes responsible for debilitating hereditary diseases. These dramatic findings, spurred by developing technology for reading genomes, have led to the Human Genome Project, which includes among its stated goals the identification of the several thousand inherited disorders recognized in man (McKusick 1991). The vast majority of disease genes recognized to date have been located using family studies where phenotypic transmission patterns have narrowed genomic positions on the basis of linkage analysis with DNA marker loci. There exist, however, a large number of more complex heritable characters that are not amenable to pedigree analysis, because of incomplete penetrance, genetic heterogeneity, or the requirement for an environmental component for phenotypic ascertainment. The latter category would include genes involving talent (requiring training), psychological alterations such as alcoholism (requiring alcohol consumption), or resistance to infectious diseases (requiring exposure). The diagnostic phenotype for such characteristics can be sporadic and refractory to phenotypic recognition in parents and offspring of affected individuals. For example, a group of hepatitis B–infected patients who develop neither hepatitis nor liver cancer in their lifetime might share a common "resistance gene" that could be scored only in relatives also exposed to hepatitis B virus (Beasley et al. 1981). This restriction is the reason why so few human loci that require an environmental component have been identified, while they certainly must exist, since several dozen such genes are identified in animal gene maps (O'Brien and Evermann 1988; Kozak 1993).

To address this difficulty as well as to increase the information content of patients with rare diseases in the absence of available family data, association analysis based on linkage disequilibrium with DNA markers has been explored theoretically (Nei and Li 1973; Lander and Botstein 1986; Chakraborty and Smouse 1988; Chakraborty and Weiss 1988; Risch 1992; Briscoe et al. 1994). Indeed, significant linkage disequilibrium has been influential in identifying mutations responsible for several human hereditary diseases, including cystic fibrosis, Huntington disease, diastrophic dysplasia, and diabetes (Todd et al. 1988; Kerem et al. 1989; Hästbacka et al. 1992; Huntington's Disease Collaborative Research Group 1993). In spite of these ad-

Received February 10, 1994; accepted for publication May 5, 1994.

Address for correspondence and reprints: Dr. Stephen J. O'Brien, Chief, Laboratory of Viral Carcinogenesis, National Cancer Institute, Frederick, MD 21702-1201.

* On secondment from Department of Biology, School of Biological Sciences, Macquarie University, Sydney, North Ryde, New South Wales 2109, Australia.

0002-9297/94/5504-0023$02.00

vances, linkage disequilibrium is not widely employed as a mapping tool, primarily because it is rather uncommon in human populations over linkage intervals >1.0 cM (Bodmer 1986). The occurrence of three major causes of linkage disequilibrium (recent mutation, founder effects, and epistatic selection) is thought to be rare, and accumulated disequilibrium decays rapidly (within a few generations) except for very tightly linked loci (Weir 1990). This means that a comprehensive human mapping screen predicated on linkage disequilibrium would require an extremely high-resolution linkage map of 3,000–4,000 markers spaced at ~1-cM intervals along the 3,200-cM human genome.

In an effort to enhance the power of linkage disequilibrium for detecting linkage, several authors (Chakraborty and Weiss 1988; Risch 1992; Briscoe et al. 1994) have developed an approach that emphasizes a fourth cause of linkage disequilibrium, recent admixture of genetically divergent populations, as a possible strategy to detect genetic linkage. The method takes advantage of the tendency of isolated populations to diverge from each other in the polymorphic allele frequencies at homologous loci. When two populations that have been isolated for a long period connect and exchange genes, the difference in allele frequencies produces appreciable linkage disequilibrium. The amount of disequilibrium generated between pairs of loci depends on the difference between allele frequencies in the two populations (δ) and does not require actual linkage, so unlinked and linked loci both produce temporary linkage or "gametic" disequilibrium. The decay of gametic disequilibrium among unlinked genes is rapid (within 2–4 generations), but disequilibrium between linked genes decays more slowly. The method of *m*apping by *a*dmixture *l*inkage *d*isequilibrium (MALD) exploits the differential of gametic disequilibrium between linked versus unlinked genes.

In order to determine actual limits for the feasibility of MALD in available human populations, we first identify the parameters that would influence the extent and persistence of admixture linkage disequilibrium (ALD) in recently admixed populations such as have occurred because of human migration to the New World within the past few centuries (Briscoe et al. 1994). The effect of these parameters (listed in table 1) on linkage ascertainment is modeled here by analytic and computer simulations that approximate realistic situations of human genetic analysis and disease cohorts. The simulations demonstrate the feasibility of MALD in detecting linked genes in the absence of family data and quantify the influence of (1) θ, the recombination distance between markers and disease locus; (2) m, the fraction of genetic admixture; (3) g, generation time since admixture; (4) δ, the difference in homologous allele frequencies in admixed populations; and (5) N, sample size of patient cohorts that is required to resolve linkage. The simulation results provide practical guidelines for mapping genes by using affected patients from recently admixed populations (3–15 generations since initial admixture) and by selecting loci, at 10–20-cM intervals, that have maximal divergence between the initial populations.

Simulation Model Overview

Two general categories of simulations were performed. The first group, models Ia and Ib, takes a genomewide approach to admixed populations that undergo random matings for 15 generations. The models presume a simple scenario of admixture in which genetically differentiated populations fuse with random matings thereafter. Ia and Ib are largely deterministic and illustrate the dynamics of linkage disequilibrium across a genome over 15 generations. The second simulation set, models IIa and IIb, examines the potential for linkage disequilibrium ascertainment over a 60-cM chromosomal segment in an admixed population of 5,000 individuals. Model IIa tests for the effects of θ, g, N, and δ on the detectability of ALD over background (table 1). Both categories of simulations were designed to approximate actual situations of patient cohorts, available genetic markers, and human genome structure. We also address in model IIb the effect of continuous admixture between parental populations over several generations, a situation more reflective of human population admixture natural histories.

Linkage Disequilibrium over the Genome after Equal Admixture of Genetically Differentiated Populations: Model I

The first model considers the pattern of differential persistence of ALD following fusion and gene flow between two previously isolated, panmictic, and genetically differentiated populations of equivalent size. The initial level of linkage disequilibrium, D_0, in a population formed by admixture between two genetically differentiated populations, depends primarily on the allele frequency differences between the founding populations (Chakraborty and Smouse 1988; Briscoe et al. 1994):

$$D_0 = m(1-m)\delta_A\delta_B, \quad (1)$$

where δ_A is the difference in allele frequency at locus A, δ_B is the difference in allele frequency at locus B, and m and $(1-m)$ are the fractional contributions of the two founding populations. In this formulation, we have assumed that the loci are in linkage equilibrium in each founding population. Subsequent levels of disequilibrium would depend on the recombination frequency between loci, the number of generations after the initial admixture, and the level of further genetic contact with the founding populations.

The assumptions and parameters (fixed and variable) prescribed by model Ia are listed in table 1 and are as follows: First, 100 diallelic codominant autosomal loci were distributed randomly among 22 autosomes of total genetic

Table 1

Parameters That Influence the Amount, Persistence, and Detectability of Linkage Disequilibrium in an Admixed Population: Simulation Values in Text Models

	Model I		Model II	
	a	*b*	*a*	*b*
Genome parameters:				
No. of autosomes	22	22	...	...
Relative chromosome size	Per human	Per human	...	...
Length of genome (cM)	2,500	2,500	...	...
No. of loci typed	100	100	5	5
Recombination distance, (θ)	0–50	0–50	0–40[a]	0–40[a]
Population parameters:				
Population size	Infinite	Infinite	5,000	5,000
F_x of admixture at generation i, m	$m_0 = .5; m_i = 0$	$m_0 = .5; m_i = 0$	$m_0 = .5; m_i = 0$	$m_0 = \ldots = m_7 = \alpha = .03, .05, .07;$ $m_8 = m_9 = 0$
No. of generations since admix, g	0–15	0–15	0–9	2–9
No. of individuals sampled, N	Deterministic	Deterministic	50, 100, 200, and 300	50, 100, 200, and 300
Genetic parameters:				
Transmission modality	Codominant	Codominant	Codominant	Codominant
Allele frequency	0–1.0	0–1.0	0–1.0	0–1.0
$\lvert\delta\rvert = \lvert f(A_I) - f(A_{II})\rvert$	0–.2	.2–1.0	.1, .2, .3, and .4	.1, .2, .3, and .4
Simulation parameters:				
No. of simulation replicas	1	1	1,000	1,000
Stringency of significance	...	...	.05 and .01	.05 and .01
Method of disequilibrium	D′	D′	χ^2	χ^2

[a] θ values in simulation II include B-X distances of 1, 2, 3, ... 10 cM (see fig. 3), thereby covering 19, 18, 17, ... 10 cM for the X-C distance, 21, 22, 23, ... 30 cM for the A-X distance, and 39, 38, ... 30 cM for the X-D distance. The combination of X and the other four loci therefore spans 0–40 cM, in 1-cM units (see text).

length 2,500 cM. Chromosome lengths are proportionate to human autosomes, as calculated elsewhere (Stephens et al. 1990). The average spacing between consecutive loci is somewhat greater than 25 cM because of the division of the genome into 22 autosomes.

Second, the allele frequencies in one founding population, e.g., $f(A_I)$, were uniform random variables chosen from 0–1.0. Third, the magnitude of δ for each locus, $|\delta| = |f(A_I) - f(A_{II})|$, was chosen as a uniform random variable from 0–.2. Note that the assumption of a fixed upper limit for $|\delta|$ reflects that the founding populations have some maximum level of divergence for the test loci.

Fourth, the gene frequency in the second founding population—here, $f(A_{II})$—was chosen by randomly adding or subtracting $|\delta|$ from the other founder's frequency, $f(A_{II}) = f(A_I) \pm |\delta|$. If this produced a frequency outside [0,1], then the opposite was performed. For instance, if $f(A_I) = .9$ and $|\delta| = .2$ and had addition been chosen, leading to $f(A_{II}) = 1.1$, subtraction would be substituted, $f(A_{II}) = .7$. In this way the actual δ values were equally likely to be positive or negative. However, for convenience we will refer to δ values in all that follows as if they were strictly positive. This convention follows from the symmetry of diallelic systems: we can always choose the allele whose frequency is higher in population I.

Finally, the admixed population was initiated by equal input (m=.5) from the two founding populations, with random mating within the admixed population for 15 generations thereafter. This model is meant to approximate a human genome admixture RFLP screen with markers spaced, on the average, 25 cM apart, with a rather conservative limit of $\delta \leq .2$.

Linkage disequilibrium for all pairwise comparisons among the 100 loci was computed for each generation of the simulation. Both D, the conventional measure of linkage disequilibrium,

$$D = f_{AB} - f_A f_B, \tag{2}$$

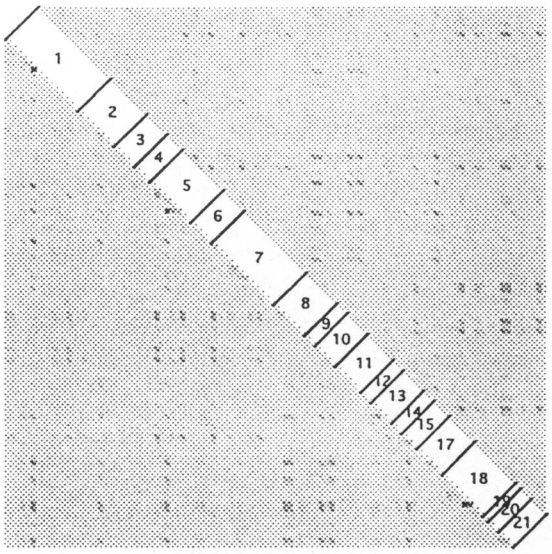
and D', the absolute value of Lewontin's (1974) measure of proportionate disequilibrium,

$$D' = |D/D_{max}|, \tag{3}$$

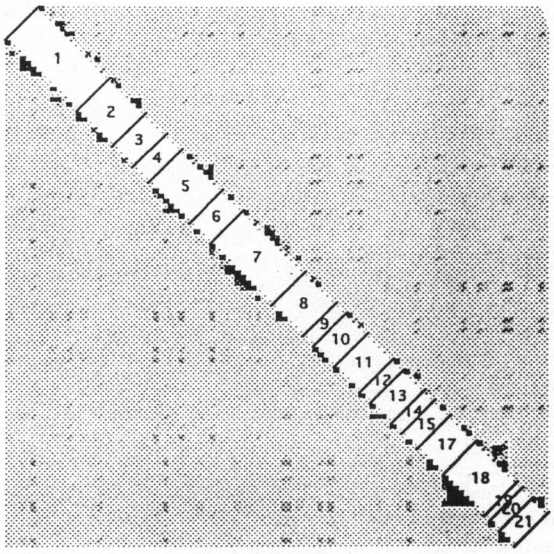
were computed. In equation (2), f_{AB} is the frequency of the AB gamete, and f_A and f_B represent the frequencies of alleles A and B, respectively. In equation (3), D_{max} is the maximum value of D for a set of gene frequencies: min $[f_A(1-f_B), f_B(1-f_A)]$ if $D > 0$, and $-\min[f_A f_B, (1-f_A)(1-f_B)]$ if $D < 0$. Because of the influence of initial allele frequency

Pairwise D' values, $\delta \leq 0.20$

Generation 0



Generation 6

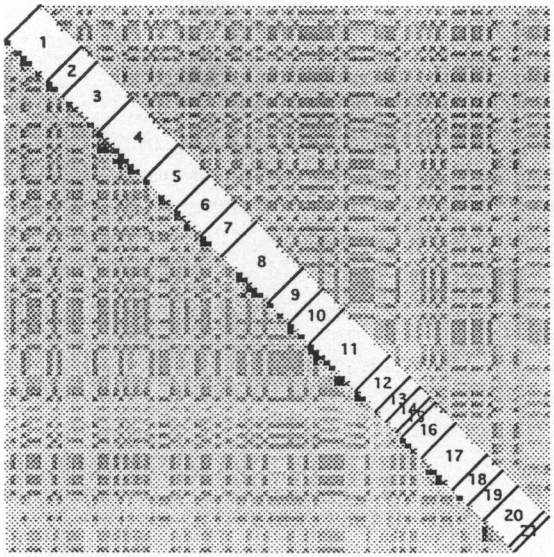


Generation 3

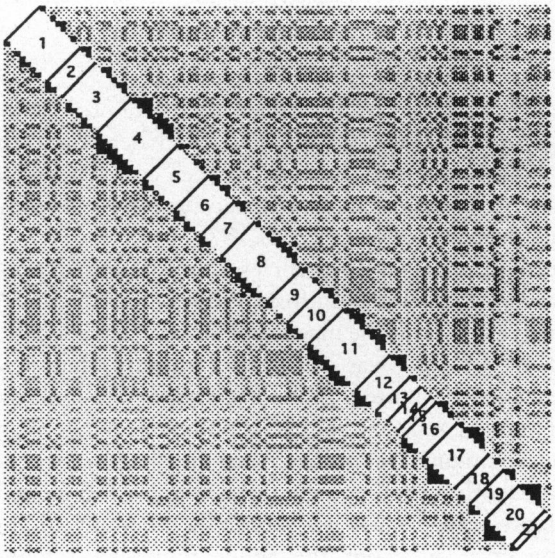
Generation 9

Pairwise D' values, $\delta \geq 0.20$

Generation 0



Generation 6



Generation 3

Generation 9

G0	G3	G6	G9
D' ≤ 0.25	D' ≤ 0.031	D' ≤ 0.0039	D' ≤ 0.0005
0.25 < D' ≤ 0.50	0.031 < D' ≤ 0.063	0.0039 < D' ≤ 0.0078	0.0005 < D' ≤ 0.0010
0.50 < D' ≤ 0.75	0.063 < D' ≤ 0.094	0.0078 < D' ≤ 0.0117	0.0010 < D' ≤ 0.0015
0.75 < D' ≤ 1.00	0.094 < D' ≤ 0.125	0.0117 < D' ≤ 0.0156	0.0015 < D' ≤ 0.0020
—	0.281 < D'	0.0351 < D'	0.0044 < D'

on D values, we used D' to evaluate disequilibrium between linked versus unlinked genes across the genome.

Of the 4,950 pairwise comparisons for 100 loci dispersed over the 2,500-cM genome, 162 pairs were linked (i.e., $\theta<.50$), and 4,788 were not. The extent of gametic disequilibrium, estimated by D' at generations 0, 3, 6, and 9 is illustrated from five *representative simulations* in figure 1. Loci have been put into their known order, starting with the most telomeric locus on 1p, proceeding through each autosome to the last locus on chromosome 22. Hence, comparisons between linked loci fall on or near the diagonal.

In generation 0, gametic disequilibrium is evident between both unlinked and linked loci. In subsequent generations the disequilibrium between unlinked genes ("background") will be reduced by ½ each generation (Chakraborty and Weiss 1988; Briscoe et al. 1994). By generation 6 the highest D' values occur primarily between linked genes; by generation 9 the differential is even more apparent. In table 2, we examined the efficiency of D' to identify linked marker pairs within discrete centimorgan intervals in the context of the model Ia simulation. By generation 6, 46% (50/109) of markers separated by 0–30 cM were detectable above background, and by the 10th generation 80% of markers <30 cM apart and 91% of markers <20 cM apart were detected by a D' value above the $D'_{max-unl}$ produced by unlinked markers ($D'_{max-unl}$; table 2). The results of model Ia (upper panels of fig. 1; table 2) illustrate the trend, over the entire genome, for the persistence of linkage disequilibrium of loosely linked genes (recall that the average distance between markers is 25 cM), relative to the background of unlinked genes. The differential is pronounced after about six generations following admixture between two populations whose gene markers are only modestly differentiated ($\delta\leq.2$).

In model Ib, we intentionally increased our efficiency for ascertainment of D' (and thereby linkage) by increasing δ values between the admixing populations by using an exponentially decreasing distribution from .2 to 1.0. In all other respects, models Ia and Ib are similar. The results of the simulations are illustrated in the lower panels of figure 1 and are tabulated in table 3. As expected (eq. [1]), the initial D' values are higher than with lower δ values employed by model Ia (compare generation 0; upper vs. lower panels of fig. 1); however, the elevated D' values (relative to $D'_{max-unl}$) accumulate rapidly along the diagonal in subsequent generations, identifying disequilibrium between linked genes over the background of unlinked locus pairs (lower panels of fig. 1). Many pairs of linked genes are detectable by generation 3, and all pairs of linked loci with $\theta \leq .30$ are detectable by seven generations after admixture (table 3). Even loci with $.3 < \theta \leq .4$ are detectable by 10 generations. Thus we conclude that choosing loci with large δ values greatly enhances the efficiency of detection, at least for the current simplistic scenario.

The simple model of admixture in model I is the same as that explored by Chakraborty and Smouse (1988) and Chakraborty and Weiss (1988). Our detection criterion, that D' be $>D'_{max-unl}$ at some generation after the initial admixture event, is amenable to an explicit solution. First, consider that as long as $D'_0 > 0$ and $\theta < .5$, linkage would eventually be detected in this idealistic scenario. We can solve for the generation at which linkage becomes detectable as $g_{det} = -\ln(D'_0)/\ln[2(1-\theta)]$. Thus, detectability depends entirely on D'_0 and θ. In turn, D'_0 depends on δ values and the allele frequencies. D' values will always be larger than the corresponding D values but generally will not be very high unless the allele frequencies are noncentral and disparate ($D>0$) or similar ($D<0$). In the left panel of figure 2, we show D'_0 values corresponding to $\delta_A = \delta_B = .10$ and the full range of allele frequencies for both loci. In the right panel of figure 2, we show the upper limit of θ that would be detectable by a certain generation, as a function of D'_0. Detectability rises sharply as D'_0 increases, such that even loose linkage ($\theta>.20$) would be detectable in three to nine generations.

Model I: Using Polymorphic Human Loci

In the paper that accompanies this one (Dean et al. 1994 [in this issue]), we report the allele frequencies in four human ethnic groups (Caucasian, African American, Chinese, and American Indian) for 257 polymorphic DNA markers. The linkage relationship of each of these loci was estimated from the human framework map (NIH/CEPH Collaborative Mapping Group 1992) or from their cytogenetic position relative to adjacent loci included on the human linkage map (Dean et al. 1994). In all there were 479 locus pairs that were linked at ≤30 cM and 189 pairs at ≤10 cM. The distribution of δ values between Caucasians and African Americans

Figure 1 Amount of linkage disequilibrium (estimated by D') between pairwise combinations of 100 loci dispersed on 22 autosomes of a 2,500-cM genome (model I). The upper two panels show model Ia, $\delta \leq .2$. Simulation parameters are listed in table 1. Typed loci are arranged in a linear array along the 22 autosomes as delineated by chromosome numbers along the diagonal. In generation 0 there is considerable disequilibrium between unlinked loci because of admixture, dictated entirely by δ values of the loci (eq. [1]), and there is no tendency for larger D' values to fall near the diagonal. Disequilibrium between unlinked loci decays by 50% each generation, while the disequilibrium between linked genes persists (Chakraborty and Weiss 1988; Briscoe et al. 1994). This conclusion is evident by the concentration of higher D' values at or near the diagonal, by generation 6. Note that the scale of D' decreases with increasing generation. In this way, the pattern of "background" values for unlinked genes is frozen in contrast to that of linked genes. As are shown in the lower two panels, simulation conditions for model Ib are identical to those of Ia, except $.2 \leq \delta \leq 1.0$ (see text).

Table 2

Detectability of Linkage as a Function of Generations Elapsed since Admixture ($\delta \leq .20$)

	No. Detected as $D' > D'_{max-unl}$ when θ =					Detection Criterion
	.0–.1	.1–.2	.2–.3	.3–.4	.4–.5	$D' > D'_{max-unl}$[a]
Generation:						
0	0	0	0	0	0	1.000000
1	0	0	0	0	0	.500000
2	0	0	0	1	0	.250000
3	3	0	0	1	0	.125000
4	9	4	2	1	0	.062500
5	13	12	5	2	0	.031250
6	24	19	7	3	0	.015625
7	30	25	15	4	0	.007813
8	33	28	16	5	0	.003906
9	34	31	16	8	0	.001953
10	36	35	16	8	0	.000977
11	37	35	18	9	0	.000488
12	38	36	21	11	0	.000244
13	38	38	22	12	0	.000122
14	38	38	25	14	1	.000061
15	38	39	27	17	1	.000031
No. of linked pairs	38	40	31	30	23	. . .

[a] Theoretical maximum D′ for unlinked genes. For unlinked genes, D′ decays by ½ each generation, so D′ values above this maximum would indicate linkage of the two markers.

Table 3

Detectability of Linkage as a Function of Generations Elapsed since Admixture ($\delta \geq .20$)

	No. Detected as $D' > D'_{max-unl}$ when θ =					Detection Criterion
	.0–.1	.1–.2	.2–.3	.3–.4	.4–.5	$D' > D'_{max-unl}$[a]
Generation:						
0	0	0	0	0	0	1.000000
1	12	5	3	6	7	.500000
2	20	9	7	10	7	.250000
3	33	20	10	13	7	.125000
4	37	26	21	16	7	.062500
5	38	29	24	21	8	.031250
6	38	29	26	21	8	.015625
7	38	29	27	23	8	.007813
8	38	29	27	27	9	.003906
9	38	29	27	29	12	.001953
10	38	29	27	30	15	.000977
11	38	29	27	30	18	.000488
12	38	29	27	30	19	.000244
13	38	29	27	30	19	.000122
14	38	29	27	30	19	.000061
15	38	29	27	30	19	.000031
No. of linked pairs	38	29	27	30	27	. . .

[a] Theoretical maximum D′ for unlinked genes. For unlinked genes, D′ decays by ½ each generation, so D′ values above this maximum would indicate linkage of the two markers.

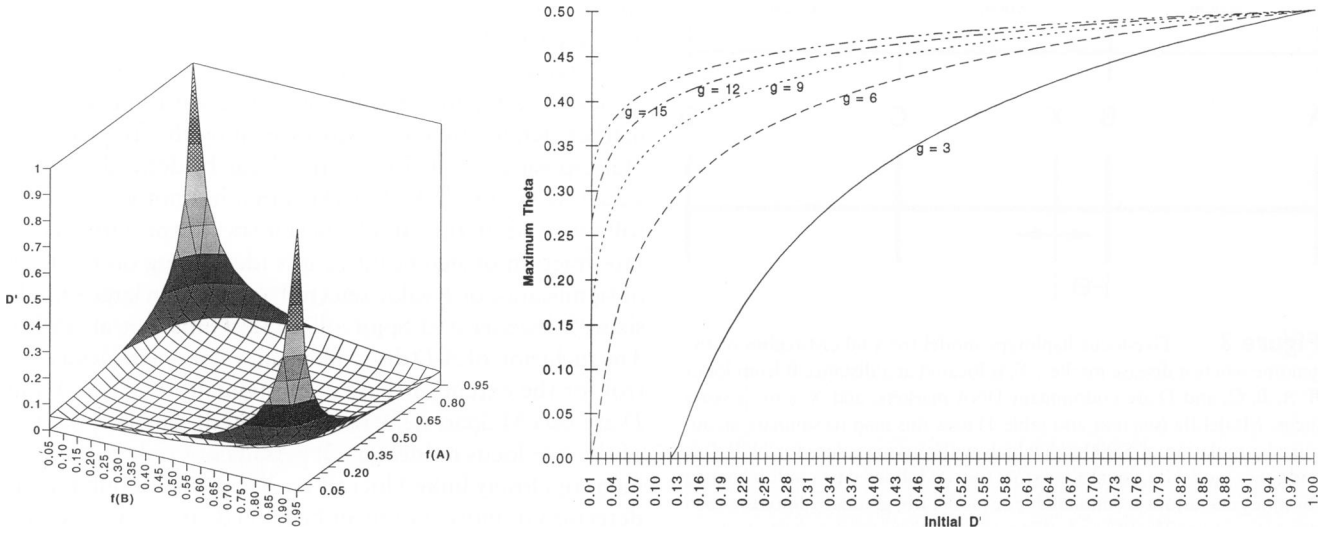


Figure 2 *Left panel,* Range of possible D' values when $\delta = .10$, as a function of allele frequencies (in the admixed population) at locus *A* and locus *B*. *Right panel,* Maximum θ for which linkage will be detectable by generation g, as a function of D'_0. Curves for $g = 3, 6, 9, 12$, and 15 are shown. Admixture scenario of model I parameters is assumed.

revealed that 36% (93 loci) had $\delta \geq .2$. (This value is likely an underestimate of African-Caucasian δ, since our sample included only African Americans.)

We used these actual human values of f, δ, and θ to predict the detectability of these linkage relationships by MALD, assuming the admixture scenario of model I (table 4). The results, which are comparable to the theoretical values of tables 2 and 3, show that we would detect 254 (53%) of the 479 pairs linked at ≤30 cM by generation 6. Of 338 pairs linked at ≤20 cM, 216 pairs (64%) would show elevated linkage disequilibrium in the African American population by generation 6.

Table 4

Detectability of Linkage as a Function of Generations Elapsed since Admixture

	No. Detected when θ =					Detection Criterion
	.0–.1	.1–.2	.2–.3	.3–.4	.4–.5	$D' > D'_{max-unl}$
Generation:						
0	0	0	0	0	0	1.000000
1	0	0	0	0	0	.500000
2	5	6	0	0	0	.250000
3	37	14	3	1	0	.125000
4	77	33	6	1	0	.062500
5	110	53	18	4	0	.031250
6	140	76	38	16	0	.015625
7	161	95	57	26	0	.007813
8	172	111	76	38	0	.003906
9	180	124	88	48	0	.001953
10	184	132	97	57	0	.000977
11	188	136	102	70	0	.000488
12	188	139	105	77	0	.000244
13	188	139	112	88	2	.000122
14	189	144	115	95	6	.000061
15	189	147	118	102	7	.000031
No. of linked pairs	189	149	141	161	96	. . .

Note.—Data are based on actual human loci with $\theta < .50$ and measured δ between Caucasian and African American population samples (Dean et al. 1994).

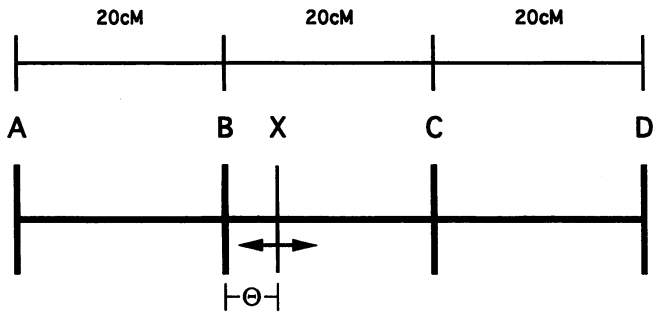


Figure 3 Five-locus haplotype model for a 60-cM region of the genome where a disease marker, *X*, is located at a distance θ from locus *B*. *A*, *B*, *C*, and *D* are codominant DNA markers, and *X* is the disease locus. Model IIa (see text and table 1) uses this map to simulate an admixed population of 5,000 individuals with varying values for *N*, θ, *f*, δ, and *g*. For model IIa, equal admixture (*m*=.5) followed by random mating for nine generations is simulated. For model IIb, gradual admixture each generation (*m*=.03, .05, and .07, per generation) until two generations prior to the sampled generation is assumed (see text).

Model II: Measuring Empirical Limits for θ, δ, g, and N by Using MALD

To evaluate the influence of genetic and population parameters on the efficiency of MALD to detect a disease locus, we employ the five-locus haplotype model illustrated in figure 3. In this model four polymorphic diallelic RFLP marker loci–*A*, *B*, *C*, and *D*–are separated by 20-cM intervals without interference, and the disease marker, *X*, is located between marker *B* and *C* at a set distance θ (range 1–10 cM) from locus *B*. For each simulation fixed values for θ, δ, *g*, and *N* are assigned (table 1 lists the actual values) and tested using a population admixture model in the following manner. For model IIa, an initial population, at $g = 0$, of 5,000 individuals (10,000 gametes) was derived from an equal admixture (*m*=.5) of two populations, followed by random mating for nine generations. The allele frequencies of the five loci in the founding populations were determined by a uniform random distribution 0–1.0, as in model I, and all loci were considered to be in linkage equilibrium in the founding populations. The δ between the two populations was set at a specific value for all five loci in each simulation, as was the sample size, *N*, and the *B*-*X* distance, θ (table 1). For each simulation the initial five-locus haplotype frequencies were determined; a sample of individuals was selected; the population was replaced under the assumption of random mating with recombination; and the population was sampled again. The process was repeated for every generation (up to nine), and the simulation was replicated 1,000 times under each set of parameter combinations (δ, *N*, and θ; see table 1). The statistical significance of disequilibrium between all pairs of loci was monitored by calculating $\chi^2 = ND^2/[f_A(1-f_A)f_B(1-f_B)]$, where f_A and f_B are the frequency of alleles at two test loci, in this case, *A* and *B*. The χ^2 value has an approximate χ^2 distribution with 1 df for diallelic loci (Weir 1990).

An important goal of our analysis was to test the efficiency of detecting true linkage of *X* with linked DNA markers while simultaneously estimating the frequency of "false-positives." "False-positives" can be derived from residual ALD of unlinked markers that has not yet decayed following admixture and from statistical expectations of a rare fraction of significant results (depending on the level of significance or *P* value selected) detected in large sample sizes (Schweder and Spjotvoll 1982; Gilbert et al. 1990). The inclusion of *A*-*D* disequilibrium values provides a control for the extent of these in all simulations, since *A* and *D* are 60 cM apart (effectively unlinked). Further, the use of the five-locus model (fig. 3) permits not only the assessment of closely linked loci (*B* vs. *X*) but also simultaneous detection of more loosely linked loci (*C* to *X*, *A* to *X*, and *D* to *X*), providing a complete picture of *X*'s linkage at θ values of 1–39 cM. Finally, an important advantage to the five-locus simulation strategy is that it allows us to measure the concordant linkage disequilibria of two or more loci with the disease locus, because *A*, *B*, *C*, and *D* are all linked to *X*, albeit at different θ values (fig. 3).

To illustrate the rapid decay of "gametic disequilibrium" between unlinked genes (*A*-*D*) relative to linkage disequilibrium for linked loci in admixed populations, we present in figure 4 simulation results for $\delta = .4$, $\theta = 10$ cM from *B*, $N = 300$, and $m = .5$. Since δ and *N* are fairly large, conditions are favorable for the detection of *X*'s linkage to *B*, and, indeed, the frequency of detection is relatively high for at least five generations after the initial admixture event. However, for at least two generations after the initial admixture event, the level of false-positives would also be unacceptably high, as judged by the number of times that markers *A* and *D* are in significant disequilibrium with each other. At three generations or beyond, the level of false-positives drops to ≤20%, and we see that the more stringent criterion ($P<.01$) dramatically reduces this level relative to the more relaxed criterion ($P<.05$). A summary of the incidence of *A*-*D* disequilibrium (for unlinked loci) detected in simulations over different ranges of δ, *N*, and *P* is presented in table 5. For empirical study design, these results (fig. 4; table 5) emphasize that the first few generations after admixture would be too soon to use MALD for linkage detection, because of nonspecific "gametic disequilibrium" between unlinked loci.

Our five-locus strategy allows us to examine simultaneously linkage disequilibria for four loci, each with different θ intervals from *X*. In the top panel of figure 5, we show the distribution of specific outcomes for the 1,000 replications of the conditions used in figure 4 (δ=.4; N=300; g=3; θ=10 cM). That is, the top panel of figure 5 shows the number of times, in 1,000 simulations, that significant ALD with *X* was detected for specific loci (*A*, *B*, *C*, or *D*), specific locus pairs, specific triplets, or all four loci simultaneously. The highest

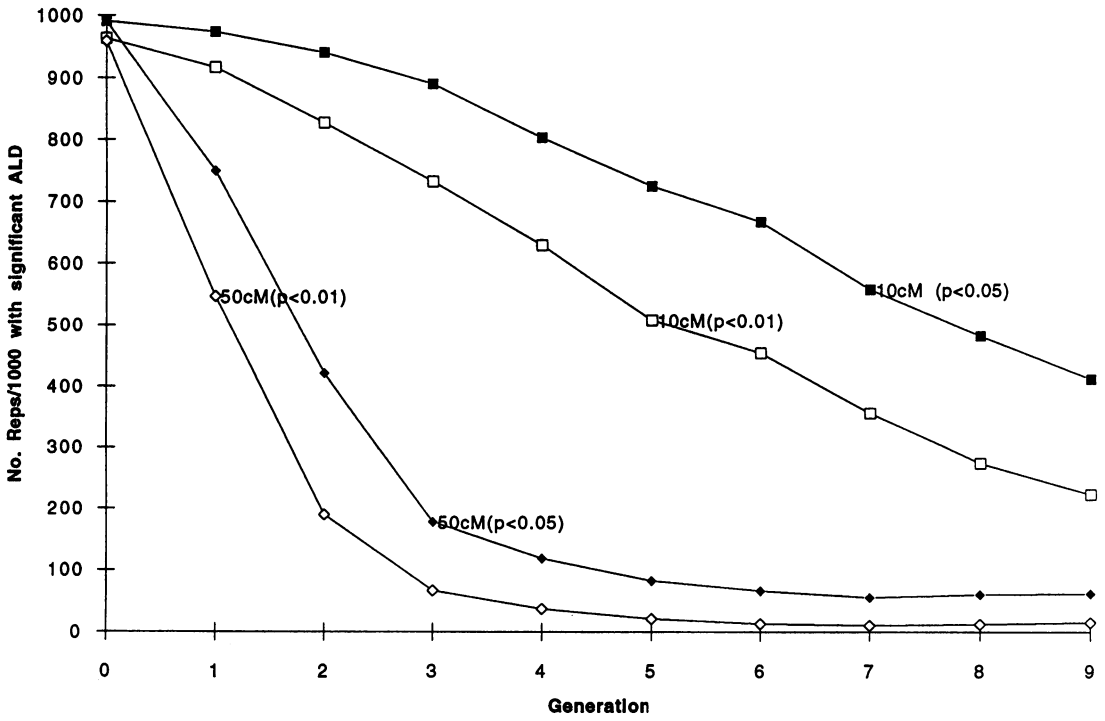


Figure 4 Number of simulation replications (of 1,000) with significant ALD. Two curves (labeled "10cM") show this value for a *B-X* distance of θ = 10 cM, as a function of *g* for high ($P \leq .01$) and low ($P \leq .05$) statistical stringency. Likewise, two curves (labeled "50cM") show this value (between unlinked markers *A* and *D*) as a function of *g*, for high and low stringency. Model IIa simulation conditions with δ = .4 and *N* = 300 were used.

values were evident with *B* or *C*, and the concordant ALD with both *B* and *C* was only 10% lower than single-locus ALD. This result emphasizes the power of MALD to identify two linked marker loci with a disease locus as an indicator of linkage. Increasing the statistical stringency (to $P \leq .01$) results in a decrease in detected disequilibria across all categories, but the categories involving *A* or *D* are dropped more dramatically than other combinations.

In fact, it may be more useful to focus on the genomic interval, rather than on the specific markers. Noting that all four markers are, in fact, linked to *X*, we can ask whether at least one, two, three, or all four markers at once are in disequilibrium with *X*. These results are illustrated in the bottom panel of figure 5. We see that the criterion requiring at least one of the four markers to be in significant ($P \leq .05$) disequilibrium with *X* detected linkage

Table 5

Frequency (%) of False-Positives (Significant Linkage Disequilibrium between Markers A and D)

	Low Stringency ($P<.05$) for *N* =				High Stringency ($P<.01$) for *N* =			
	50	100	200	300	50	100	200	300
Generation 3:								
δ = .2	5.37	5.44	6.63	7.31	1.14	.98	1.46	1.92
δ = .3	5.87	6.83	9.16	10.96	1.10	1.78	2.49	3.55
δ = .4	6.74	9.69	14.62	18.93	1.70	2.78	4.55	6.84
Generation 6:								
δ = .2	5.03	5.35	5.88	5.97	.87	1.09	1.37	1.37
δ = .3	5.64	5.62	6.35	6.46	1.16	1.13	1.38	1.58
δ = .4	5.06	5.38	6.77	6.55	1.07	1.02	1.74	1.55
Generation 9:								
δ = .2	5.38	5.55	5.72	6.16	1.01	1.33	1.30	1.40
δ = .3	5.21	5.69	6.14	6.27	1.22	.92	1.37	1.26
δ = .4	5.26	5.29	5.84	5.82	1.03	.96	1.37	1.30

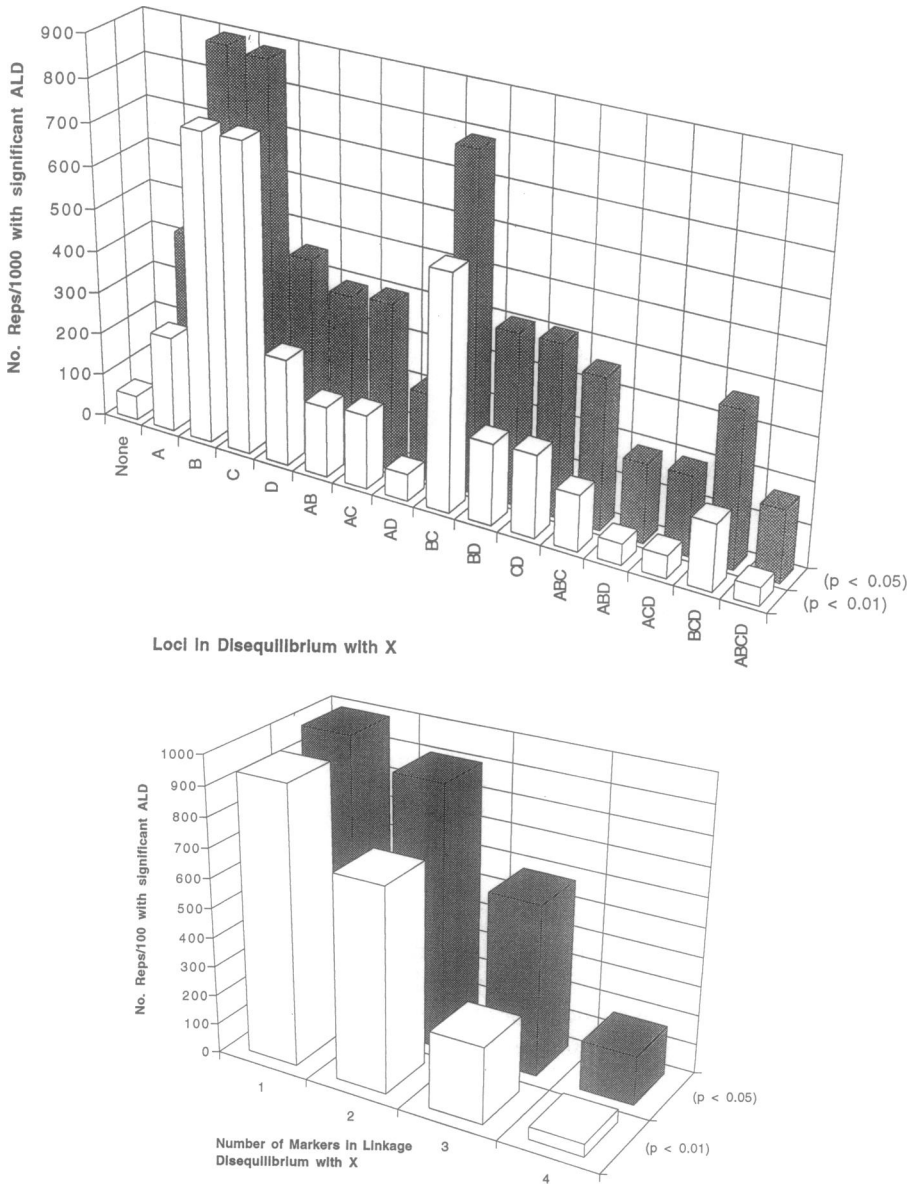


Figure 5 Number of simulation replications (of 1,000) with significant ALD under different detection strategies. Model IIa simulation conditions with $\delta = .4$, $N = 300$, $g = 3$, and the *B-X* distance $\theta = 10$ cM were used. *Top panel,* Replications in which no markers; at least markers *A, B, C,* or *D;* at least *A and B, A and C,* . . . ; or all four markers were in statistically significant disequilibrium with *X*. Results under two statistical stringency levels are shown. Note that categories are not all mutually exclusive (e.g., category *A* includes *ABCD* and may overlap with other categories). *Bottom panel,* As in the top panel, but without identifying a specific marker from the *A-D* region. Hence, categories are at least one, at least two, at least three, or all four markers in significant ALD with *X*.

to *X* in >95% of the simulations. Furthermore, increasing the stringency ($P \leq .01$; 94.3%) or requiring disequilibrium (low stringency) with any two markers (90.4%) worked almost as well. We have shown (fig. 4) that an increased stringency will cut down on the number of false-positives, and we expect that the demand for concordance of two or more markers will likewise reduce this number. Perhaps disappointingly, it does not appear that demanding increased concordance (three or more markers simultaneously) would be an efficient criterion for detecting linkage, at least under the map density modeled in figure 3.

The results of several selected sets of simulation conditions are illustrated in figure 6. We plot the number of simulation runs (of 1,000) for which significant linkage disequilibrium (at high [$P \leq .01$] and relaxed [$P \leq .05$] statistical stringency) is detected between *X* and the four linked markers, in figure 3, as a function of θ; the latter varies, for each simulation, by 1 cM, at increasing intervals from the *B* marker locus. The results of a rather ideal (and optimal for MALD) scenario in the upper-left panel of figure 6 consider a population of 5,000 at generation 3, after equal admixture ($m = .5$), where $\delta = .4$ for each locus, the sample

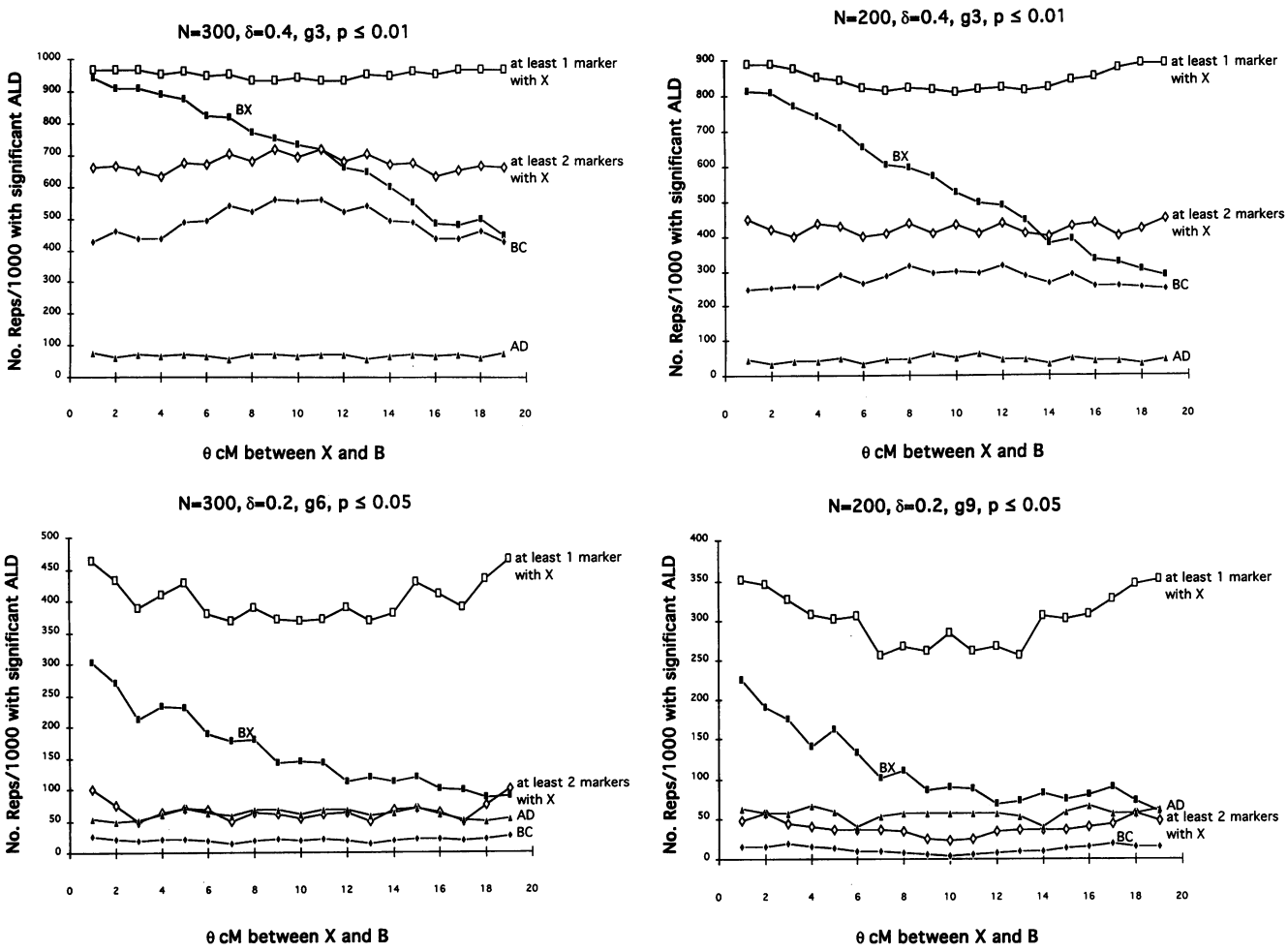


Figure 6 Simulation results under four representative combinations of simulation conditions. Results for four tests (*BX*=at least marker *B* in significant ALD with *X;* and *BC*=both *B* and *C* in significant ALD with *X*) are shown. In addition, the *AD* curve shows the number of times the "unlinked" markers *A* and *D* were in significant ALD. *Upper-left panel,* $N = 300$; $\delta = .40$; $g = 3$; $P \leq .01$. *Upper-right panel,* $N = 200$; $\delta = .40$, $g = 3$, $P \leq .01$. *Lower-left panel,* $N = 300$; $\delta = .20$; $g = 6$; $P \leq .05$. *Lower-right panel,* $N = 200$; $\delta = .20$; $g = 9$; $P \leq .05$.

size $N = 300$ patients, and the stringency is high ($P \leq .01$). In this example, detected linkage disequilibrium between the disease locus, *X*, and the *B* marker, ranges from 93% to 48% as a function of θ, while unlinked genes (*A-D*) display disequilibrium 6%–8% of the time. If we ask how frequently one or more of the linked markers are in disequilibrium with *X*, the value approaches 98% and, importantly, is largely independent of θ. This means that MALD under ideal conditions ($\delta \geq .4$; $N \geq 300$; $g \geq 3$) is close to 100% effective in linkage ascertainment with markers spaced 20 cM apart, even at high stringency ($P \leq .01$). The two loci closest to *X* (*B* and *C*) are *both* in linkage disequilibrium with *X* 42%–56% of the time. At least two of the four linked markers simultaneously display linkage disequilibrium with *X* 65%–70% of the time and again appear unaffected by the actual position of *X* between *B* and *C*.

We have simulated a variety of different starting parameters of δ, *N*, *g*, θ, and *P*, most of which show less efficiency than under the ideal conditions, in the upper-left panel of figure 6. In all, we examined 2,880 sets of simulation conditions, (4 *N* values, 4 δ values, 10 θ values, 9 generations after admixture, and 2 *P* values; see table 1). In the upper-right panel of figure 6, the starting parameters were identical to those in the upper-left panel, except sample size, $N = 200$ patients. In this case the detection of *B-X* linkage decreases from 81% to 30% as the *B-X* distance increases from 1 to 19 cM. Nevertheless, with a sample of 200 affected individuals, 88%–90% of the time, at least one marker displays disequilibrium, and two markers are in disequilibrium 40%–45% of the time, regardless of the position of *X* relative to *B* and *C*. In fact, both *B* and *C* would simultaneously be in disequilibrium ~25%–30% of the time, compared with ~3%–4% of unlinked genes (*A-D*) at the high level of statistical stringency ($P \leq .01$).

In all the simulations, the parameter that had the largest effect on ascertainment was δ between marker and disease

Table 6

Efficiency of ALD (% Detection) to Map Genes in Populations

	No. of Patients	1 Locus in ALD with X		2 Loci in ALD with X
		$P \leq .05$	$P \leq .01$	($P \leq .05$)
$\delta = .4$:				
G3	100	78–86	49–59	38–44
G3	200	95–98	81–89	73–76
G3	300	99–100	93–97	89–92
G9	100	38–72	13–44	7–12
G9	200	56–90	28–75	16–22
G9	300	71–97	41–89	23–28
$\delta = .2$:				
G3	100	26–30	6–9	2–4
G3	200	36–41	12–15	6–8
G3	300	45–50	18–23	9–13
G9	100	21–29	5–7	1–3
G9	200	26–35	7–13	2–6
G9	300	31–42	10–17	4–8
$\delta = .4$; $\alpha = .05$:				
G9	100	59–75	27–48	17–24
G9	200	81–92	53–77	40–44
G9	300	91–98	71–92	59–63

loci, as would be expected from equation (1). For example, in the lower-left panel of figure 6, when $\delta = .2$, $N = 300$, $g = 6$, and $P \leq .05$, the *B-X* pair is detected in MALD only 30% of the time when the *B-X* distance is 1 cM, and this drops to 10% at 19 cM. Nevertheless, at least one of the four markers is significantly in disequilibrium 40%–47% of the time, and again this result is largely independent of the position of *X* relative to *B* or *C*. A lower limit of the MALD parameters is presented in the lower-right panel of figure 6. In this simulation (δ=.2; N=200; g=9; and $P \leq .05$), 25%–35% of the time at least one marker locus shows disequilibrium with *X*, and this is under relaxed statistical stringency ($P \leq .05$).

Our results emphasize the utility of two primary test criteria for detecting linkage disequilibrium of disease gene *X* with the genomic interval represented by loci *A*, *B*, *C*, and *D* (fig. 3): (1) high stringency ($P \leq .01$) significance of the disequilibrium between *X* and at least one marker; and (2) low stringency ($P \leq .05$) significance of the disequilibrium between *X* and two or more markers. In table 6 and figure 7, we summarize all the results for the entire range of parameters, on the basis of these two test criteria. As expected, the efficiency of detection improves as either sample size or δ increases. This conclusion holds as long as $g \geq 3$, and for both test criteria shown in figure 7. The results shown in figure 6 suggest that both of these test criteria would be relatively insensitive to the actual θ value, except for test 1 in late generations (fig. 6, lower-right panel). For both tests, the efficiency of detection drops as the number of generations increases, although more severely for the two-locus test than for the single-locus test.

Continuous Infusion of Genes in an Admixed Population: Model IIb

We have considered so far an idealistic model of admixture that is suited to an experimental setting but that has only limited relevance to natural populations. Many species, including humans, are not amenable to an experimental manipulation of this sort but may nonetheless have a history of recent admixture that may be favorable for gene mapping studies. The admixture model used in our simulations can easily be generalized by the following modifications. First, consider that at generation 0, a parent population receives a fraction (m_o) of its genomes from an introgressing population. In our previous examples, $m_o = .5$ in generation 0, and $m_i = 0$ for subsequent generations. In model IIb, we set m_i to some specified value (α) for each generation so that the sampled population at generation g has its original genetic composition reduced to $(1-\alpha)^g$.

For various human groups, particularly in the New World, admixture has occurred in the past several hundred years, or 10–20 generations. Current estimates suggest that ~30% of the gene pool of African Americans is of European origin (Chakraborty 1986). We have used a simple scenario of admixture in which $m_0 = m_1 = \ldots = m_7 = \alpha$ and $m_8 = m_9 = 0$, to model a hypothetical sample from the African American gene pool. In this model, it is assumed that European haplotypes were introduced into the African American gene pool at a rate of α per generation for the first eight generations (~120–160 years). We know from models I–IIa that linkage disequilibrium between unlinked genes will compromise efforts to detect linked genes, especially within the first two or three generations after admixture. Therefore, we have restricted our hypothetical sample to individuals who have no Caucasian parents or grandparents (m_8=m_9=0), a condition that would be incorporated in MALD patient cohorts.

The steady rate of introduction of Caucasian genes during the first eight generations suggests that the remaining proportion of African genes will be $(1-\alpha)^8$. In model IIb, we have tested three values of α (.03, .05, and .07) corresponding to values of 22%, 34%, and 44%, respectively, for the frequency of genes of European origin in the African American gene pool. We assumed that a map of marker loci exists, each locus with δ of .40, and that the gene being mapped (*X*) also has a δ of .40. Under this model, all sampling occurs at generation 9.

Figure 8 and table 6 show results obtained for model IIb, tests 1 and 2. Both tests have reasonable efficiencies of detection as long as the sample size is $\geq$200, with great improvement as sample size increases. The considerable overlap among the three curves suggests that there is very little effect of the precise α value, at least compared with the effect of sample size. The stringency of ascertainment

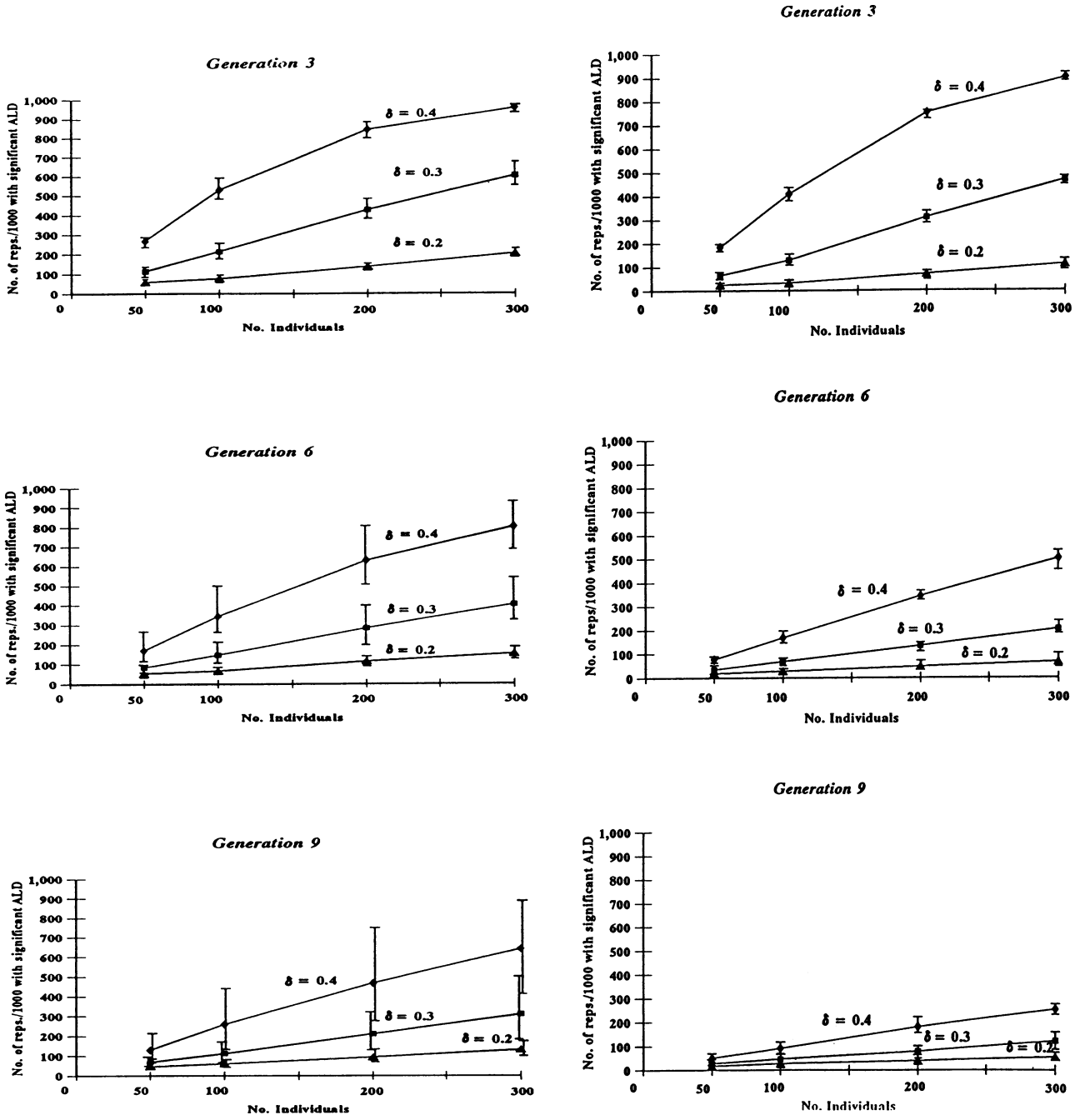


Figure 7 Summary of results under model IIa. *Left panel,* Results under test 1: at least one marker in significant ALD with *X*, under high stringency ($P \leq .01$). *Right panel,* Results under test 2: at least two markers in significant ALD with *X*, under low stringency ($P \leq .05$). Error bars indicate the observed range of values as θ varies from 1 to 10 cM.

has a marked effect in this scenario, because at low stringency there is appreciable "false-positive" *A-D* linkage disequilibrium (table 7). These results imply that MALD has appreciable power to detect linkage over 20-cM intervals in recently admixed human populations specifically selected to exclude patients with parents or grandparents from the introgressing population.

Discussion

The phenomenon of linkage disequilibrium in natural populations has been studied in depth both theoretically and empirically (Lewontin and Kojima 1960; O'Brien and MacIntyre 1971; Nei and Li 1973; Brown 1975; Clegg 1984; Chakraborty and Smouse 1988; Chakraborty and

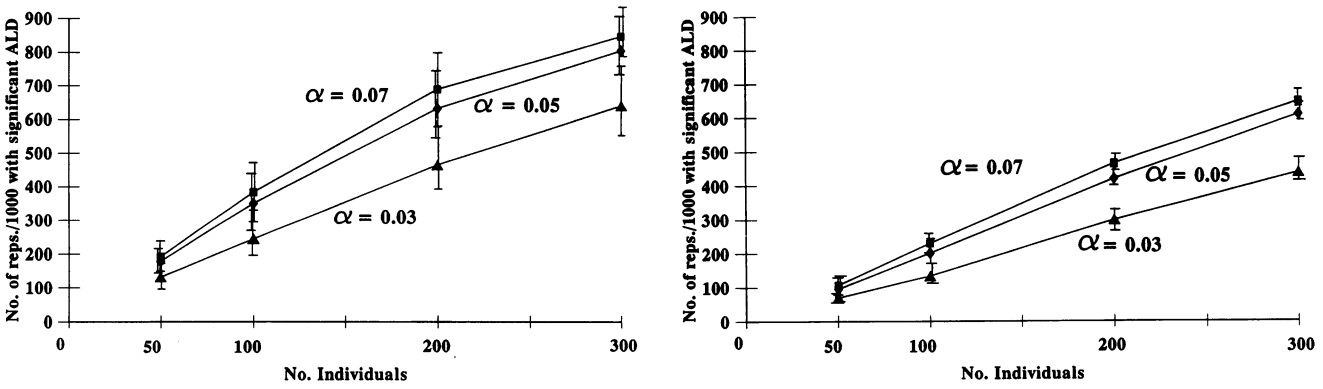


Figure 8 Summary of results under model IIb. Values for α are the admixture frequencies maintained over generations 0–7 of the simulation. *Left panel,* Results under test 1 for generation 9: at least one marker in significant ALD with *X*, under high stringency ($P \leq .01$). *Right panel,* Results under test 2 for generation 9: at least two markers in significant ALD with *X*, under low stringency ($P \leq .05$). Error bars indicate the observed range of values as θ varies from 1 to 10 cM.

Weiss 1988; Weir 1990). Linkage disequilibrium originates largely from one of four events in the natural history of a population: (1) recent mutation, (2) founder effects, (3) epistatic selection, and (4) admixture of genetically differentiated populations. Linkage disequilibrium caused by the first two factors tends to decay rapidly in panmictic populations, except for that between tightly linked loci (≤1 cM). We note that between tightly linked genes linkage disequilibrium has been a very powerful approach for identifying human disease genes by using DNA markers (Todd et al. 1988; Kerem et al. 1989; Hästbacka et al. 1992; Huntington's Disease Collaborative Research Group 1993). Several authors have made theoretical arguments suggesting that recent admixture of genetically differentiated populations (such as racial admixture of human populations in the New World in the past 4 centuries by European explorers) would result in more extensive linkage disequilibrium over large (≤20 cM) genomic intervals. We explore here the influence of genetic, genomic, and population parameters (table 1) that affect the power of MALD, by using computer simulations designed to approximate human populations and screens for disease loci, particularly those requiring an environmental component for phenotypic ascertainment (e.g., resistance to infectious agents). Our results provide empirical guidelines for design of a genomewide disease-locus screen in admixed human populations by using available DNA polymorphisms that have divergent allele frequencies in different human populations (Dean et al. 1994 [in this issue]).

When gene flow resumes between genetically differentiated populations, linkage disequilibrium between pairs of homologous loci with different allele frequencies is produced as a function of δ (eq. [1]), regardless of whether the locus pairs are linked (Nei and Li 1973; Chakraborty and Smouse 1988; Chakraborty and Weiss 1988; Briscoe et al. 1994). We show here that "gametic disequilibrium" between unlinked loci decays in 2–4 generations, while that between genes linked at distances ≤20 cM decays more slowly. The detection of the differential persistence of linkage disequilibrium between linked versus unlinked locus pairs forms the basis for linkage ascertainment using MALD. To minimize the effects of "false-positives" due to both gametic disequilibrium between unlinked genes and occurrence of a few statistically significant departures from expectation that would occur in large samples, we impose two test criteria for an ascertainment: (1) set statistical stringency at $P \leq .01$ and (2) track the incidence of at least two adjacent loci in disequilibrium with the disease locus *X*. In practice, disequilibria between a disease locus *X* and markers throughout the genome would be esti-

Table 7

Frequency (%) of False-Positives in Model IIb (Significant Linkage Disequilibrium between Markers A and D)

	Low Stringency ($P \leq .05$) for N =				High Stringency ($P \leq .01$) for N =			
α	50	100	200	300	50	100	200	300
.03	6.45	6.76	8.32	9.56	1.34	1.37	2.26	2.89
.05	5.78	7.17	7.97	10.46	1.12	1.52	1.97	3.32
.07	6.25	7.33	8.81	11.06	1.28	1.59	2.38	3.29

mated; among all significant values, the best place to look would be the marker(s) with the highest significance and/or genomic regions where clusters of markers are in significant disequilibrium with X.

As suggested by equation (1), δ, the difference in allele frequency between populations, has a profound effect, not only on the initiation of disequilibrium, but on its detectability (figs. 6 and 7). This is true regardless of generation tested, sample size taken, or test used for detection. Thus, loci selected for MALD should be as differentiated as possible between the parent populations. In fact, when δ falls below .2, these markers (regardless of their heterozygosity) are virtually uninformative, with detectability seldom reaching 20% in the best of circumstances (fig. 7). The maintenance of disequilibrium, and hence its detectability, is controlled by three additional factors: recombination (θ), number of generations after admixture (g), and whether disequilibrium is restored by continued contact with one of the parental populations ($m_i>0$). In early generations, the distance from the disease locus X and flanking markers (θ) is inconsequential, but by the ninth generation following admixture (in the simple scenario of equal admixture at generation 0, followed by panmixis and no further immigration), the θ value exerts a marked effect on the detection of linkage disequilibrium (see "error bars" in the left panel of fig. 7).

An increased sample size will also improve detectability under most circumstances (fig. 7). However, there may be a concomitant gain in false-positives (tables 5 and 7), especially if δ is large and g is small. There does not appear to be a tendency for the frequency of false-positives to increase with increases in either δ or sample size in later generations. Although enhanced detectability with increasing sample size is reasonably general, increasing the sample size cannot easily compensate for markers with low δ values (fig. 7; δ=.2).

A simulation model with a more realistic (for human populations) scheme of continuous gene immigration over several generations (fig. 8) did not appreciably alter any of these conclusions. With the empirical adjustment to include only patients with parents and grandparents who were themselves descendants of admixture (e.g., for African Americans, exclude those with Caucasian parents and grandparents), we strove to eliminate most of the "false-positive" linkage disequilibrium of unlinked genes that decays in two to four generations (fig. 4). With continuous gradual admixture, we still see a profound effect of δ, an increased ascertainment with increasing sample size, a modest effect of θ in the context of our 20-cM marker interval model (fig. 3), and a slight effect of the actual per-generation value of α (fig. 8). The level of false-positives shows an increase with increasing sample size but bears no relationship with α (table 7). This increase is reduced to 2%–3% when high statistical stringency ($P\leq.01$) is imposed (table 7) and in any case is much lower than is seen for generation 3 of the equal admixture ($m_0=.5$; $m_i=0$) model IIa (see the upper-left panel of fig. 6). The simulations of continuous admixture are admittedly simplistic in the context of actual human histories, but the relative insensitivity of MALD to vacillations in m_i or θ levels gives hope to the prospect of detecting marked disequilibrium between linked genes in selected human populations.

From the presented results, we can now recommend some limits on empirical patient cohort design for efficacious MALD. For an ideal situation, we suggest a collection of equally spaced markers at 10–20-cM intervals throughout the human genome. The markers should have maximum differentiation (δ) between the parent populations, ideally $\delta \geq .4$. In the report that accompanies this one, we list the distribution of 257 polymorphic markers dispersed, on the average, at 10–20-cM intervals (Dean et al. 1994 [in this issue]). The modal δ value is .15 between Caucasians and African Americans. This value is presumably an underestimate of the actual δ between Caucasians and Africans (by ~30%), since we have measured the modern African American population rather than the ancestral African population from which African Americans are descended. Although an estimated $\delta = .2$ is lower than ideal, there are >7,000 human polymorphic loci that could be screened in humans; a collection of markers approaching $\delta = .4$ could certainly be achieved (also see Roychoudhury and Nei 1988).

As listed in table 6, a cohort group of 300 patients whose history included admixture in the past 15 generations (continuous or not) but not in the past few generations would be the best candidates for MALD. Under conditions that show higher δ or smaller θ, fewer patients would be required, as illustrated in table 6, where (for δ=.4; N=200) the detection of linkage is 95%–98% in generation 3 and 56%–90% by generation 9 when the lower-stringency criterion ($P\leq.05$) is used.

The models presented here are simple but emphasize the power of using a human map of diallelic DNA markers with MALD for detecting disease loci. There are certain limitations, however, that can reduce the power of MALD that merit discussion. First, all computations here assumed codominant expression of marker and disease phenotype. The effect of dominant and recessive modalities for disease expression diminishes ascertainment in a manner that requires further analysis (J. C. Stephens, unpublished information). Second, quantitative traits that affect many hereditary diseases will produce reduced signals but may be discernable under certain conditions. Third, the allele frequencies for the disease gene must also be different (δ_x) between the admixing populations for detection, and this will vary between traits. Finally, the development of new high-resolution polymorphisms such as microsatellite polymorphism and SSCP adds enormous power to family studies but will have a special effect on MALD, because of multiple alleles and high mutation rates (Litt and Luty

1989; Weber and May 1989; Poduslo et al. 1991; Weber and Wong 1993). Each of these factors will have a measurable effect on MALD ascertainment, and their influence will be the object of future analysis (J. C. Stephens, unpublished information).

Acknowledgments

We would like to thank Drs. M. T. Clegg, B. S. Weir, and M. Dean for their comments as this work was in progress.

References

Beasley RP, Hwang LY, Lin CC, Chien CS (1981) Hepatocellular carcinoma and hepatitis B virus: a prospective study of 22,707 men in Taiwan. Lancet 2:1129–1133

Bodmer WF (1986) Human genetics: the molecular challenge. In: Quantitative biology. Vol. 1. Cold Spring Harb Symp Quant Biol 51:1–13

Briscoe D, Stephens JC, O'Brien SJ (1994) Linkage disequilibrium in admixed populations: applications in gene mapping. J Hered 85:59–63

Brown AHD (1975) Sample sizes required to detect linkage disequilibrium between two or three loci. Theor Popul Biol 8: 184–201

Chakraborty R (1986) Gene admixture in human populations: models and predictions. Yearbook Phys Anthropol 29:1–43

Chakraborty R, Smouse P (1988) Recombination in haplotypes leads to biased estimates of admixture proportions in human populations. Proc Natl Acad Sci USA 85:3071–3074

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

Clegg MT (1984) Dynamics of multi-locus genetic systems. Oxf Surv Evol Biol 1:160–183

Dean M, Stephens JC, Winkler C, Lomb D, Ramsburg M, Boaze R, Stewart C, et al (1994) Polymorphic admixture typing in human ethnic populations. Am J Hum Genet 55:788–808 (in this issue)

Gilbert DA, Reid YA, Gail MH, Pee D, White C, Hay RJ, O'Brien SJ (1990) Application of DNA fingerprints for cell-line individualization. Am J Hum Genet 47:499–514

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nature Genet 2:204–211

Huntington's Disease Collaborative Research Group, The (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72: 971–983

Kerem BS, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, et al (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Kozak CA (1993) Retroviral and cancer related genes of the mouse. In: O'Brien SJ (ed) Genetic maps, locus maps of complex genomes, 6th ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, p 4.143–4.156

Lander ES, Botstein D (1986) Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 51:49–62

Lewontin RC (1974) The genetic basis of evolutionary change. Columbia University Press, New York

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397–401

McKusick VA (1991) Mendelian inheritance in man, 6th ed. Johns Hopkins University Press, Baltimore

Nei M, Li WH (1973) Linkage disequilibrium in subdivided populations. Genetics 75:213–219

NIH/CEPH Collaborative Mapping Group (1992) A comprehensive genetic linkage map of the human genome. Science 258:67–86

O'Brien SJ, Evermann JF (1988) Interactive influence of infectious disease and genetic diversity in natural populations. Trends Ecol Evol 3:254–259

O'Brien SJ, MacIntyre RJ (1971) Empirical demonstration of a transient linkage disequilibrium in *Drosophila*. Nature 230: 335–336

Poduslo SE, Dean M, Kolch U, O'Brien SJ (1991) Detecting high-resolution polymorphisms in human coding loci by combining PCR and single-strand conformation polymorphism (SSCP) analysis. Am J Hum Genet 49:106–111

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

Roychoudhury AK, Nei M (1988) Human polymorphic genes: world distribution. Oxford University Press, New York

Schweder T, Spjotvoll E (1982) Plots of P-values to evaluate many tests simultaneously. Biometrika 69:493–502

Stephens JC, Cavanaugh ML, Gradie MI, Mador ML, Kidd KK (1990) Mapping the human genome: current status. Science 250:237–244

Todd JA, Bell JI, McDevitt HO (1988) HLA antigen and insulin-dependent diabetes. Nature 333:710

Weber JL, May PE (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388–396

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128

Weir BS (1990) Genetic data analysis: methods for discrete population genetic data. Sinauer, Sunderland, MA