Joint Linkage and Linkage Disequilibrium Mapping in Natural Populations

Rongling Wu* and Zhao-Bang Zeng†

**Department of Statistics, University of Florida, Gainesville, Florida 32611 and †Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695*

Manuscript received April 19, 2000
Accepted for publication October 30, 2000

ABSTRACT

A new strategy for studying the genome structure and organization of natural populations is proposed on the basis of a combined analysis of linkage and linkage disequilibrium using known polymorphic markers. This strategy exploits a random sample drawn from a panmictic natural population and the open-pollinated progeny of the sample. It is established on the principle of gene transmission from the parental to progeny generation during which the linkage between different markers is broken down due to meiotic recombination. The strategy has power to simultaneously capture the information about the linkage of the markers (as measured by recombination fraction) and the degree of their linkage disequilibrium created at a historic time. Simulation studies indicate that the statistical method implemented by the Fisher-scoring algorithm can provide accurate and precise estimates for the allele frequencies, recombination fractions, and linkage disequilibria between different markers. The strategy has great implications for constructing a dense linkage disequilibrium map that can facilitate the identification and positional cloning of the genes underlying both simple and complex traits.

WITH improved techniques for high-throughput identification and genotyping of polymorphisms, it has been possible to genotype molecular markers throughout the genome and construct a dense linkage map covering the entire genome (Landegren *et al.* 1998; Wang *et al.* 1998). For species with homozygous inbred lines available, the linkage analysis of markers is based on the recombinations of a particular chromosome region that are created by hybridization between two genetically divergent inbred lines (Mather and Jinks 1982). Such a strategy can directly provide an estimate for the linkage relationship of markers as measured by recombination fraction, because there is clear information about parental linkage phase between alleles of different markers. However, linkage mapping has two major limitations. First, for closely linked markers, there will be few recombinations in a segregating generation and, hence, a dense linkage map will provide little extra information about the localization of target genes, unless the number of individuals of the generation is very large (Darvasi *et al.* 1993). For example, using linkage analysis, Long *et al.* (1995) could only map quantitative trait loci (QTL) affecting bristle numbers in Drosophila to regions of ~5–10 cM. Second, homozygous inbred lines used to generate the F_1 parents of *a priori* known linkage phases for the traditional linkage analysis (Mather and Jinks 1982) are virtually unavailable for natural populations. For many study materials sampled randomly from a natural population, therefore, uncertainty of linkage phase between markers prevents a direct estimate of their recombination fraction.

For natural populations, the degree of nonrandom genetic association or linkage disequilibrium, produced at a historic time by various evolutionary forces such as mutation, drift, selection, and admixture, is estimated to indirectly infer how strongly these markers are linked on the same chromosome. If the linkage disequilibrium of the markers occurred a long time ago, a strong linkage disequilibrium detected may suggest close physical linkage between the markers because linkage disequilibrium decays with time (Kaplan *et al.* 1995). This principle has tremendous potential for constructing fine-scale linkage disequilibrium maps for cloning the genes that cause complex qualitative or quantitative traits (reviewed in Templeton 1999). At present, a number of theories or techniques have been well established for linkage disequilibrium-based mapping of target genes in natural populations (Hästbacka *et al.* 1992, 1994; Risch and Merikangas 1996; Xiong and Guo 1997; Terwilliger and Weiss 1998; Kruglyak 1999; Meuwissen and Goddard 2000).

The success of linkage disequilibrium mapping is the presence of linkage disequilibrium between different loci arising from the covariance of the population frequencies of nonalleles in the same gamete (Lynch and Walsh 1998). The degree and extent of linkage disequilibrium reflect the evolutionary history of a population and its interactions with different evolutionary forces (Hill and Robertson 1968; Nei and Li 1973; Ohta 1982a,b; Epperson and Allard 1987; Peterson *et al.* 1999; Farnir *et al.* 2000). A number of earlier studies

Corresponding author: Rongling Wu, Department of Statistics, 533 McCarty Hall C, University of Florida, Gainesville, FL 32611. E-mail: rwu@stat.ufl.edu

have focused on statistical methods for detecting the existence of linkage disequilibrium. A likelihood-based procedure was developed by Hill (1974) to estimate the coefficient of linkage disequilibrium between two loci in a finite random mating population. Brown (1975) established a theoretical framework for the sample sizes required for detecting the disequilibrium by the use of data on gametic and zygotic frequencies. Weir and Cockerham (1978) suggested a statistical procedure for calculating the power of testing linkage disequilibrium between different loci of multiple alleles. More recently, Luo (1998) and Luo and Suhai (1999) proposed statistical approaches for testing and estimating linkage disequilibrium between a polymorphic marker and a putative QTL. All of these analyses have laid a necessary foundation for linkage disequilibrium mapping of disease genes in human populations (Hästbacka *et al.* 1992, 1994; Collins and Morton 1998; Escamilla *et al.* 1999; Service *et al.* 1999).

A major problem with current strategies for linkage disequilibrium mapping is that they provide little insight into the mechanistic basis of linkage disequilibrium detected in a natural population. Without such knowledge, however, the genomic localization and cloning of genes based on linkage disequilibrium may not be successful, because a strong linkage disequilibrium detected between two genetic loci may be due to the recent occurrence of disequilibrium rather than a close physical map distance of the two loci. In human genetics, the cause of linkage disequilibrium can be revealed through a combined linkage and linkage disequilibrium analysis, as shown by a transmission/disequilibrium testing (TDT) approach (Allison 1997; Rabinowitz 1997; Camp 1998). However, TDT is critically relied upon for nuclear family data with complete records for multiple successive generations. This approach therefore cannot be used for genome mapping in many other situations where no nuclear family records are available or for other undomesticated species, such as wildlife and forest trees. For these situations or species, it is essential for developing a powerful approach that needs no nuclear family but can still provide a simultaneous estimate for linkage and linkage disequilibrium between genetic loci of interest.

In this article, we propose a new strategy for detecting linkage and linkage disequilibrium between polymorphic markers in natural populations. The new strategy is expected to provide a new avenue for studying the evolutionary dynamics of population variation and differentiation. Furthermore, as compared to a pure linkage analysis or linkage disequilibrium analysis, the combined use of linkage and linkage disequilibrium analysis methods can greatly enhance the feasibility of high-resolution mapping of genes of interest and their subsequent genetic manipulation. The strategy is presented in two parts, one on dioecious species and the other on monoecious species. Dioecious species including animals, humans, and many forest trees, such as Ginkgo, poplar, and willow, display a single sex for an individual and, therefore, are predominantly outcrossing. Monoecious species comprising most crop and horticultural plants and forest trees such as pine, fir, and spruce carry both sexes on every individual and could be both self-compatible and outcrossing. We first deal with a simpler dioecious model. A more complicated statistical model for analyzing natural populations of monoecious species will be reported in a forthcoming companion article.

TWO-LOCUS MODEL

Population structure theory: Consider a panmictic natural population of a dioecious species in Hardy-Weinberg equilibrium. In the population, η neutral codominant markers $\mathbf{M}^1, \ldots, \mathbf{M}^\eta$ are assumed to be segregating. Let an allele at marker $\mathbf{M}^i$ ($i = 1, \ldots, \eta$), designated by M^i_r ($r = 1, \ldots, n_i$), have population frequency P^i_r, $\Sigma_{r=1}^{n_i} P^i_r = 1$, with the number of alleles n_i at the marker being arbitrary.

Assume that a second marker $\mathbf{M}^j$ is located on the same chromosome as $\mathbf{M}^i$, both markers having a recombination fraction θ^{ij}. These two linked markers are genetically associated in the population with the coefficient of gametic linkage disequilibrium between a pair of nonalleles from the two markers denoted by D^{ij}_{rs} ($s = 1, \ldots, n_j$). The population frequency of the gamete (haplotype) at the two markers $M^i_r M^j_s$ can be expressed as

$$P^{ij}_{rs} = P^i_r P^j_s + D^{ij}_{rs}, \tag{1}$$

with the constraints of

$$-p^i_r p^j_s \le D^{ij}_{rs} \le p^i_r(1 - p^j_s) \quad \text{(Lewontin 1964)}$$

$$\sum_{r=1}^{n_i} D^{ij}_{rs} = \sum_{s=1}^{n_j} D^{ij}_{rs} = 0 \quad \text{(Weir and Cockerham 1978).}$$

The value of D^{ij}_{rs} may be positive or negative depending on whether nonalleles M^i_r and M^j_s are in coupling ($M^i_r M^j_s$ gametes are overrepresented) or repulsion ($M^i_r M^j_s$ gametes are underrepresented) disequilibrium (Lynch and Walsh 1998). Because random mating is assumed in the population, $n_i n_j$ two-locus gametes unite to form $\frac{1}{4} n_i n_j (n_i + 1)(n_j + 1)$ unique zygotic genotypes, designated by $M^i_{r_1} M^i_{r_2} M^j_{s_1} M^j_{s_2}$, where r_1, r_2 ($r_1 \le r_2 = 1, \ldots, n_i$) and s_1, s_2 ($s_1 \le s_2 = 1, \ldots, n_j$) are the two alleles of zygotic genotype at markers $\mathbf{M}^i$ and $\mathbf{M}^j$, respectively. The genotype frequency of $M^i_{r_1} M^i_{r_2} M^j_{s_1} M^j_{s_2}$ in the current population is expressed as

$$P^{ij}_{r_1 r_2 s_1 s_2} = w(P^{ij}_{r_1 s_1} P^{ij}_{r_2 s_2} + P^{ij}_{r_1 s_2} P^{ij}_{r_2 s_1}), \tag{2}$$

where w is the indicator variable relating the marker genotypes to their frequencies,

$$w = \begin{cases} \frac{1}{2} & \text{if } r_1 = r_2 \text{ and } s_1 = s_2 \\ 1 & \text{if } r_1 = r_2 \text{ and } s_1 \ne s_2 \text{ or } r_1 \ne r_2 \text{ and } s_1 = s_2 \\ 2 & \text{if } r_1 \ne r_2 \text{ and } s_1 \ne s_2. \end{cases}$$

TABLE 1

The frequencies of different genotypes at two biallelic markers $\mathbf{M}^i$ and $\mathbf{M}^j$ in a natural population, the numbers (H) of genotypes drawn randomly from the population, and the conditional probabilities of the gamete genotypes (haplotypes) given each sampled plant of a particular marker genotype

Genotype	Sampled plant		Gamete (haplotype) produced by each sampled plant			
	Frequency	Number	$M_1^iM_1^j$	$M_1^iM_2^j$	$M_2^iM_1^j$	$M_2^iM_2^j$
$M_1^iM_1^iM_1^iM_1^j$	$P_{1111}^{ij} = (P_{11}^{ij})^2$	H_{1111}^{ij}	1	0	0	0
$M_1^iM_1^iM_1^iM_2^j$	$P_{1112}^{ij} = 2P_{11}^{ij}P_{12}^{ij}$	H_{1112}^{ij}	1/2	1/2	0	0
$M_1^iM_1^iM_2^iM_2^j$	$P_{1122}^{ij} = (P_{12}^{ij})^2$	H_{1122}^{ij}	0	1	0	0
$M_1^iM_2^iM_1^iM_1^j$	$P_{1211}^{ij} = 2P_{11}^{ij}P_{21}^{ij}$	H_{1211}^{ij}	1/2	0	1/2	0
$M_1^iM_2^iM_1^iM_2^j$	$P_{1212}^{ij} = 2(P_{11}^{ij}P_{22}^{ij} + P_{12}^{ij}P_{21}^{ij})$	H_{1212}^{ij}	$\frac{P_{11}^{ij}P_{22}^{ij} - \theta^{ij}D_{rs}^{ij}}{P_{1212}^{ij}}$	$\frac{P_{12}^{ij}P_{21}^{ij} + \theta^{ij}D_{rs}^{ij}}{P_{1212}^{ij}}$	$\frac{P_{12}^{ij}P_{21}^{ij} + \theta^{ij}D_{rs}^{ij}}{P_{1212}^{ij}}$	$\frac{P_{11}^{ij}P_{22}^{ij} - \theta^{ij}D_{rs}^{ij}}{P_{1212}^{ij}}$
$M_1^iM_2^iM_2^iM_2^j$	$P_{1222}^{ij} = 2P_{12}^{ij}P_{22}^{ij}$	H_{1222}^{ij}	0	1/2	0	1/2
$M_2^iM_2^iM_1^iM_1^j$	$P_{2211}^{ij} = (P_{21}^{ij})^2$	H_{2211}^{ij}	0	0	1	0
$M_2^iM_2^iM_1^iM_2^j$	$P_{2212}^{ij} = 2P_{21}^{ij}P_{22}^{ij}$	H_{2212}^{ij}	0	0	1/2	1/2
$M_2^iM_2^iM_2^iM_2^j$	$P_{2222}^{ij} = (P_{22}^{ij})^2$	H_{2222}^{ij}	0	0	0	1

$M_{r_1}^iM_{r_2}^iM_{s_1}^jM_{s_2}^j$, two-locus zygotic genotype; $M_r^iM_s^j$, two-locus gametic genotype; $P_{r_1r_2s_1s_2}^{ij}$, the frequency of two-locus zygotic genotype in the current generation; P_{rs}^{ij}, the frequency of two-locus gamete producing the current generation; θ^{ij}, the recombination fraction of the two markers; D_{rs}^{ij}, the linkage disequilibrium of the two markers.

If all zygotes can produce gametes for the next generation, there will be a total of n_in_j gametes for markers $\mathbf{M}^i$ and $\mathbf{M}^j$ at the entire population level. But different zygotic genotypes produce different types of gametes; only the genotypes heterozygous at both markers generate all types of gametes whose relative frequencies are affected by recombination fraction and linkage disequilibrium. Table 1 gives nine zygotic genotypes, their population frequencies, and the frequencies of gametes they produce for the next generation under a simpler biallelic model (see APPENDIX A for derivations). According to the population genetics theory (NAGYLAKI 1991), the amount of linkage disequilibrium between any two loci is reduced at the rate of recombination frequency after the population mates at random for one generation. The coefficient of linkage disequilibrium in the new generation is changed to be $(1 - \theta^{ij})D_{rs}^{ij}$. Thus, the gamete frequencies for haplotypes $M_r^iM_s^j$ in the new generation at the entire population level are

$$Q_{rs}^{ij} = P_r^iP_s^j + (1 - \theta^{ij})D_{rs}^{ij}. \quad (3)$$

Further, these gametes are randomly combined to generate the progeny $M_{r_1}^iM_{r_2}^iM_{s_1}^jM_{s_2}^j$, which are contained in seeds for plants. If there is no overlapping in reproduction between the parental and progeny generations, the frequencies of the genotypes at the two markers are the products of the frequencies of the corresponding gametes.

Sampling theory: A sample of H female plants is randomly selected from the population. The seeds of these sampled plants are collected and germinated into seedlings. In traditional quantitative genetics, these seedlings grown in a regular experimental design initiate a progeny test, which serves as the selection of best parents for the next generations (MCKEAND and BRIDGWATER 1998). Both the sampled plants and their progeny are genotyped for molecular markers $\mathbf{M}^1, \ldots, \mathbf{M}^\eta$, each of an arbitrary number of alleles. According to the sampling theory, every randomly selected plant from the original population should be one of the $\frac{1}{4}n_in_j(n_i + 1)(n_j + 1)$ distinguishable genotypes for the two markers $\mathbf{M}^i$ and $\mathbf{M}^j$, each genotype with a frequency of $P_{r_1r_2s_1s_2}^{ij}$ (see Equation 2) and a sample size of $H_{r_1r_2s_1s_2}^{ij}$ (Table 1). For dioecious species, offspring genotypes (contained in seeds) of a sampled plant are formed through combing its maternal gametes with paternal gametes from the pollen pool. These offspring virtually represent a half-sib relationship with the common mother and different (unknown) fathers. The relative fractions of different maternal gametes generated by a sampled plant of a particular marker genotype are given for two biallelic markers in Table 1. The frequencies of paternal gametes in the pollen pool at the entire population level are described by Equation 3. Because of different compositions of maternal gametes (Table 1), different marker genotypes of the sampled plants generate different compositions of progeny genotypes. The conditional probabilities ($Q_{r_1r_2s_1s_2}^{R_1R_2S_1S_2}$) of the progeny genotypes at markers

$\mathbf{M}^i$ and $\mathbf{M}^j$, given a mother plant, can be derived from Bayes' theorem, where the subscripts index the mother plant's genotype and the superscripts index the genotype of a progeny. As a simpler example, these conditional probabilities are given for two biallelic markers in Table 2. Similarly, we use $N^{R_1R_2S_1S_2}_{r_1r_2s_1s_2}$ to denote the number of progeny with a particular genotype collected from a sampled plant.

Estimation theory: The allele frequencies, linkage, and linkage disequilibrium for the markers $\mathbf{M}^i$ and $\mathbf{M}^j$ in the original population can be estimated using the random sample. To estimate these unknown genetic parameters associated with the two markers $\boldsymbol{\pi}^{ij} = (P^i_r, P^j_s, \theta^{ij}, D^{ij}_{rs})^T$, a two-stage hierarchical likelihood function of the marker data ($\mathbf{M}$) is formulated from the sampled plants and their half-sib families,

$$L(\mathbf{M}|\boldsymbol{\pi}^{ij}) = \prod_{r_1=1}^{n_i}\prod_{r_2=1}^{n_i}\prod_{s_1=1}^{n_j}\prod_{s_2=1}^{n_j}\prod_{\xi=1}^{H^{ij}_{r_1r_2s_1s_2}} P^{ij}_{r_1r_2s_1s_2}(\xi) \times \prod_{R_1=1}^{n_i}\prod_{R_2=1}^{n_i}\prod_{S_1=1}^{n_j}\prod_{S_2=1}^{n_j}\prod_{\zeta=1}^{N^{R_1R_2S_1S_2}_{r_1r_2s_1s_2}} Q^{R_1R_2S_1S_2}_{r_1r_2s_1s_2}(\zeta), \quad (4)$$

with the restrictions of $r_1 \le r_2$, $s_1 \le s_2$, $R_1 \le R_2$, $S_1 \le S_2$, where ξ and ζ are the ξth sampled plant and the ζth progeny of the sampled plant, respectively, and the other symbols have been defined as above and are given in Tables 1 and 2 when a biallelic model is assumed.

There have been a number of computational algorithms available to obtain maximum-likelihood estimates (MLEs) of the four unknowns. In this article, the Fisher-scoring algorithm based on iterations is employed (Edwards 1984) because it is easy to derive and also very fast. In terms of this algorithm, the estimates at the $(\tau + 1)$th iteration can be expressed by the score function vector $\mathbf{S}(\boldsymbol{\pi}^{ij})$ and Fisher information matrix $\mathbf{I}(\boldsymbol{\pi}^{ij})$ (appendix b). The values at the τth iteration are modified by adding to them the scores divided by the information, both evaluated at the τth iteration. This iteration continues until successive iterates differ by less than some specified amount. It is apparent that the appropriateness of the Fisher-scoring algorithm relies upon the condition that the information is not zero or the information matrix is nonsingular. In practice, it is always desirable to try several different starting values and to compare the likelihoods found after convergence. After obtaining the MLEs of the unknown parameters, the inverse of $\mathbf{I}(\hat{\boldsymbol{\pi}}^{ij})$ is calculated to estimate the sampling variances of $\boldsymbol{\pi}^{ij}$.

For the purpose of linkage mapping, the degree of linkage between the two markers under consideration is important and should be tested statistically. The hypotheses for testing for linkage are H_0 (free recombination), $\theta^{ij} = 0.5$ *vs.* H_1 (linkage), $\theta^{ij} \neq 0.5$. The likelihood-ratio (LR) test statistic has the form of

$$\mathrm{LR}_\theta = -2\log\left[\frac{L(\mathbf{M}|\tilde{P}^i_r, \tilde{P}^j_s, \theta^{ij} = 0.5, \tilde{D}^{ij}_{rs})}{L(\mathbf{M}|\hat{\boldsymbol{\pi}}^{ij})}\right], \quad (5a)$$

TABLE 2

The conditional probabilities of the joint genotypes at two biallelic markers $\mathbf{M}^i$ and $\mathbf{M}^j$ in the progeny population produced by the sampled plants of particular marker genotypes drawn randomly from a natural population

	Progeny									
Sampled plants	$M^i_1M^i_1M^j_1M^j_1$	$M^i_1M^i_1M^j_1M^j_2$	$M^i_1M^i_1M^j_2M^j_2$	$M^i_1M^i_2M^j_1M^j_1$	$M^i_1M^i_2M^j_1M^j_2$	$M^i_1M^i_2M^j_2M^j_2$	$M^i_2M^i_2M^j_1M^j_1$	$M^i_2M^i_2M^j_1M^j_2$	$M^i_2M^i_2M^j_2M^j_2$
$M^i_1M^i_1M^j_1M^j_1$	Q^{ij}_{11}	Q^{ij}_{12}	0	Q^{ij}_{21}	Q^{ij}_{22}	0	0	0	0
$M^i_1M^i_1M^j_1M^j_2$	$\frac{1}{2}Q^{ij}_{11}$	$\frac{1}{2}P^i_1$	$\frac{1}{2}Q^{ij}_{12}$	$\frac{1}{2}Q^{ij}_{21}$	$\frac{1}{2}P^i_2$	$\frac{1}{2}Q^{ij}_{22}$	0	0	0
$M^i_1M^i_1M^j_2M^j_2$	0	Q^{ij}_{11}	Q^{ij}_{12}	0	Q^{ij}_{21}	Q^{ij}_{22}	0	0	0
$M^i_1M^i_2M^j_1M^j_1$	$\frac{1}{2}Q^{ij}_{11}$	$\frac{1}{2}Q^{ij}_{12}$	0	$\frac{1}{2}P^j_1$	$\frac{1}{2}P^j_2$	0	$\frac{1}{2}Q^{ij}_{21}$	$\frac{1}{2}Q^{ij}_{22}$	0
$M^i_1M^i_2M^j_1M^j_2$	$(P^{ij}_{11}P^{ij}_{22} - \theta^{ij}D^{ij})Q^{ij}_{11} + (P^{ij}_{12}P^{ij}_{21} + \theta^{ij}D^{ij})Q^{ij}_{12} + (P^{ij}_{12}P^{ij}_{21} + \theta^{ij}D^{ij})Q^{ij}_{21} + (P^{ij}_{11}P^{ij}_{22} - \theta^{ij}D^{ij})Q^{ij}_{22}$			$(P^{ij}_{11}P^{ij}_{22} - \theta^{ij}D^{ij})Q^{ij}_{11} + (P^{ij}_{12}P^{ij}_{21} + \theta^{ij}D^{ij})Q^{ij}_{12} + (P^{ij}_{12}P^{ij}_{21} + \theta^{ij}D^{ij})Q^{ij}_{21} + (P^{ij}_{11}P^{ij}_{22} - \theta^{ij}D^{ij})Q^{ij}_{22}$					
$M^i_1M^i_2M^j_2M^j_2$	0	$\frac{1}{2}Q^{ij}_{11}$	$\frac{1}{2}Q^{ij}_{12}$	0	$\frac{1}{2}P^j_1$	$\frac{1}{2}P^j_2$	0	$\frac{1}{2}Q^{ij}_{21}$	$\frac{1}{2}Q^{ij}_{22}$
$M^i_2M^i_2M^j_1M^j_1$	0	0	0	Q^{ij}_{11}	Q^{ij}_{12}	0	Q^{ij}_{21}	Q^{ij}_{22}	0
$M^i_2M^i_2M^j_1M^j_2$	0	0	0	$\frac{1}{2}Q^{ij}_{11}$	$\frac{1}{2}P^i_1$	$\frac{1}{2}Q^{ij}_{12}$	$\frac{1}{2}Q^{ij}_{21}$	$\frac{1}{2}P^i_2$	$\frac{1}{2}Q^{ij}_{22}$
$M^i_2M^i_2M^j_2M^j_2$	0	0	0	0	Q^{ij}_{11}	Q^{ij}_{12}	0	Q^{ij}_{21}	Q^{ij}_{22}

$M^i_1M^i_2M^j_1M^j_2$, two-locus zygotic genotype; P^{ij}_{rs}, the frequency of two-locus gamete producing the current generation; Q^{ij}_{rs}, the frequency of two-locus gamete produced for the next generation; P^i_r, the allele frequency of marker $\mathbf{M}^i$; P^j_s, the allele frequency of marker $\mathbf{M}^j$.

where ^ and ~ denote the MLEs of the unknowns under H_1 and H_0, respectively. Linkage disequilibrium is an important population genetic parameter and its existence and degree reflect the dynamics of population evolution. The hypotheses for overall linkage disequilibrium between markers $\mathbf{M}^i$ and $\mathbf{M}^j$ can be formulated as H_0, all $D^{ij}_{rs} = 0$ *vs.* H_1, at least $D^{ij}_{rs} \neq 0$, whose LR test statistic is

$$\mathrm{LR}_D = -2\log\left[\frac{L(\mathbf{M}|\tilde{P}^i_r, \tilde{P}^j_s, \tilde{\theta}^{ij}, \hat{D}^{ij}_{rs} = 0, r = 1, \ldots n_i, s = 1, \ldots n_j)}{L(\mathbf{M}|\hat{\boldsymbol{\pi}}^{ij})}\right]. \tag{5b}$$

These test statistics (5a and 5b) are $\sim\chi^2$-distributed with 1 d.f. Alternatively, the hypothesis test about linkage disequilibrium can be based on the collapse of marker data into a few alleles. But such a treatment may change the power of the tests for linkage disequilibrium, as demonstrated in Weir and Cockerham (1978).

If the null hypothesis of (5a) is accepted, then a significant linkage disequilibrium detected by (5b) indicates that linkage disequilibrium between a pair of markers is not due to their strong linkage. In this case, results from pure linkage disequilibrium mapping (Luo 1998; Luo and Suhai 1999; Meuwissen and Goddard 2000) are ineffective for gene mapping. If nonsignificant linkage disequilibrium is detected for two linked markers, although this may be rare, the two markers are still useful for potential localization of a target gene. Thus, by testing simultaneously for the significance of linkage and linkage disequilibrium, our analytical approach increases both the effectiveness and efficiency of gene mapping in a natural population.

MARKER ORDERING

The principle for a joint linkage and linkage disequilibrium analysis of two markers can be extended to include more than two markers. This extension is based on two assumptions: (1) recombination between any two markers is independent of recombination between any other nonoverlapping two, *i.e.*, no crossover interference; and (2) linkage disequilibrium between one pair of markers is independent of disequilibrium between other pairs. When there are more than two markers, the most likely linkage order should give the highest likelihood value for a particular dataset. With the two assumptions described above, we propose a hidden Markov model to determine an optimal order for different markers (see also Lander and Green 1987).

Assume that all η codominant markers are derived from the same chromosome in a randomly mating population. We use $M^i_{r_i}$ ($r_i = 1, \ldots, n_i$) and $M^i_{r_{i1}}M^i_{r_{i2}}$ ($r_{i1} \leq r_{i2} = 1, \ldots, n_i$) to denote an allele (gamete) and genotype (zygote) from a marker $\mathbf{M}^i$, with the population frequencies $P^i_{r_i}$ and $P^i_{r_{i1}r_{i2}}$, respectively. For a particular order $\mathbf{M}^1, \ldots, \mathbf{M}^i, \ldots, \mathbf{M}^\eta$, $\theta^{i(i+1)}$ is used to denote a recombination fraction between two adjacent markers. The coefficient of linkage disequilibrium between a pair of nonalleles r_i and r_{i+1} from two adjacent markers is denoted by $D^{i(i+1)}_{r_i r_{i+1}}$. For a vector of unknowns $\boldsymbol{\pi} = \{P^i_{r_i}, \theta^{i(i+1)}, D^{i(i+1)}_{r_i r_{i+1}}\}^T$, a two-stage hierarchical likelihood function is formulated as

$$L(\mathbf{M}|\boldsymbol{\pi}) = \prod_{i=1}^{\eta}\Bigg[\prod_{r_{i1}=1}^{n_i}\prod_{r_{i2}=1}^{n_i}\prod_{r_{(i+1)1}=1}^{n_{i+1}}\prod_{r_{(i+1)2}=1}^{n_{i+1}} \times \prod_{\xi=1}^{H^{i(i+1)}_{r_{i1}r_{i2}r_{(i+1)1}r_{(i+1)2}}} P^{i(i+1)}_{r_{i1}r_{i2}r_{(i+1)1}r_{(i+1)2}}(\xi)\prod_{R_{i1}=1}^{n_i}\prod_{R_{i2}=1}^{n_i}\prod_{R_{(i+1)1}=1}^{n_{i+1}} \times \prod_{R_{(i+1)2}=1}^{n_{i+1}}\prod_{\zeta=1}^{N^{R_{i1}R_{i2}R_{(i+1)1}R_{(i+1)2}}_{r_{i1}r_{i2}r_{(i+1)1}r_{(i+1)2}}} Q^{R_{i1}R_{i2}R_{(i+1)1}R_{(i+1)2}}_{r_{i1}r_{i2}r_{(i+1)1}r_{(i+1)2}}(\zeta)\Bigg], \tag{6}$$

where there are the restrictions $r_{i1} \leq r_{i2}$, $r_{(i+1)1} \leq r_{(i+1)2}$, $R_{i1} \leq R_{i2}$, and $R_{(i+1)1} \leq R_{(i+1)2}$, and $P^{i(i+1)}_{r_{i1}r_{i2}r_{(i+1)1}r_{(i+1)2}}$ and $Q^{R_{i1}R_{i2}R_{(i+1)1}R_{(i+1)2}}_{r_{i1}r_{i2}r_{(i+1)1}r_{(i+1)2}}$ are accordingly defined by Equations 1 and 3.

Similarly, the MLEs of the unknown vector $\boldsymbol{\pi}$ can be obtained by the Fisher-scoring algorithm based on iterations (appendix b). The hypotheses for linkage and linkage disequilibrium for every two adjacent markers can be tested accordingly. Using the Markov chain model (6), we can only estimate the linkage disequilibria between two adjacent markers and ignore the estimates of disequilibria between distant markers. Such a result may be limited from a population genetic perspective, because one cannot detect all possible linkage disequilibria generated by evolutionary forces. However, this result can definitely facilitate genomic localization and cloning of genes because our objective is to use a nearest marker to manipulate a target gene of interest.

RESULTS

To demonstrate the statistical properties of the method proposed in this article, we analyze examples on the basis of simulations. In these examples, plants for seed collection are supposed to be randomly sampled from a natural population in Hardy-Weinberg equilibrium. The effects of different sampling schemes and parameter values on the estimates for unknowns are examined, respectively.

Effects of sampling schemes: Assume that the total number (1000) of the open-pollinated progeny collected from all sampled plants is fixed. Five different sampling schemes are generated by changing the number of the sampled plants (H), each of which corresponds to a half-sib family, and the size of progeny (N) generated by each sampled plant (Table 3). These five schemes represent few large families, many small families, and moderately sized families of a moderate number. Among all the strategies, the value for each of the genetic parameters P^i_r, P^j_s, θ^{ij}, and D^{ij}_{rs} for two hypothesized biallelic markers $\mathbf{M}^i$ and $\mathbf{M}^j$ is set to be equal (Table 3). The generation of the marker data for the H half-

TABLE 3

MLEs of the genetic parameters for two biallelic markers and standard errors of the MLEs (in parentheses) calculated from 100 simulation runs for different sampling schemes

Sampling scheme	M	N	$P_r^i = 0.5$	$P_s^j = 0.3$	$\theta^{ij} = 0.1$		$D_{rs}^{ij} = 0.12$	
					MLE	Power	MLE	Power
1	10	100	0.4991	0.3007	0.1008	0.95	0.1203	0.96
			(0.0012)	(0.0008)	(0.0007)		(0.0004)	
			(0.0012)	(0.0009)	(0.0009)		(0.0004)	
2	20	50	0.5004	0.2993	0.1005	0.93	0.1202	0.96
			(0.0013)	(0.0008)	(0.0007)		(0.0005)	
			(0.0014)	(0.0008)	(0.0007)		(0.0005)	
3	32	32	0.5016	0.3005	0.0999	0.92	0.1201	0.94
			(0.0012)	(0.0008)	(0.0007)		(0.0006)	
			(0.0012)	(0.0008)	(0.0008)		(0.0007)	
4	50	20	0.4994	0.3010	0.1006	0.91	0.1204	0.92
			(0.0016)	(0.0007)	(0.0008)		(0.0006)	
			(0.0017)	(0.0007)	(0.0010)		(0.0008)	
5	100	10	0.5007	0.3009	0.1007	0.89	0.1201	0.90
			(0.0011)	(0.0007)	(0.0007)		(0.0005)	
			(0.0011)	(0.0007)	(0.0008)		(0.0006)	

Standard errors presented in upper parentheses are averaged from 100 runs and those in lower parentheses are derived from the Fisher information index on the basis of a single run. P_r^i and P_s^j are the allele frequencies of markers $\mathbf{M}^i$ and $\mathbf{M}^j$, and θ^{ij} and D_{rs}^{ij} are the recombination fraction and linkage disequilibrium between the two markers.

sib families of equal size N includes the following two steps:

1. Randomly assign the nine joint genotypes at the two markers $\mathbf{M}^i$ and $\mathbf{M}^j$ to the H sampled plants according to multinomial distribution with the probabilities as given in Table 1.
2. Randomly assign the two-marker genotypes to the open-pollinated progeny generated by each sampled plant of a particular marker genotype according to the probabilities of the marker genotypes of the progeny given in Table 2.

Because the estimates for the four unknowns are based on known marker genotypes, a likelihood-based approach has many desirable properties in the rate of convergence to achieve stable MLEs and the accuracy and power to obtain these estimates (Hill 1974). In this simulation, we compare the predicted variances of the estimates for these parameters from the asymptotic variance-covariance matrix of the MLEs to the empirical estimates calculated from the repeated simulations. The estimates of the parameters based on 100 runs are averaged and their standard errors are calculated. The sampling errors of the estimates based on a single run are calculated using the Fisher information index as described in appendix b.

Table 3 illustrates the MLEs for each of the four unknown parameters and two types of standard errors under different sampling schemes. In all situations, regardless of the combinations of family number (H) and size (N), the MLEs of the allele frequencies for two hypothesized markers using the estimation procedures developed in this article are adequately consistent with their actual values. The same is also true for the MLEs of recombination fraction and linkage disequilibrium between the two markers. Results from statistical tests based on Equations 5a and 5b indicate that the alleles of these two different markers are physically significantly linked and genetically significantly associated in the population.

In this example, the predicted values for standard errors estimated from the inverse of the information matrix are reasonably approximate to their empirical values from multiple simulation runs (Table 3). This may be partly because our parameter estimates are based on complete marker information without missing data (see also Luo and Suhai 1999). As assessed by these two types of estimates for standard errors, the method proposed here has good precision for estimating the population genetic parameters of molecular markers. The power to detect significant linkage or linkage disequilibrium between the two simulated markers is higher for few families of large sizes than for many families of small sizes. But in all sampling schemes, the power is 0.90 or higher.

Effects of linkage and linkage disequilibrium: In this simulation, we assume five biallelic markers with a known order on the same chromosome. These markers are jointly sampled from a natural population in which allele frequency is set to be $P_{ri}^i = 0.40$ for each marker. The sampling strategy used is 10 half-sib families and 100 progeny in each family. Different recombination

fractions and linkage disequilibria of two adjacent markers are hypothesized as given in Table 4 and lead to four combination patterns: (1) tight linkage and weak disequilibrium, (2) tight linkage and strong disequilibrium, (3) loose linkage and weak disequilibrium, and (4) loose linkage and strong disequilibrium. We first use separate analyses for every two adjacent markers, which are then followed by a joint analysis combining all the five markers through a Markov chain model. The MLEs for unknown parameters are obtained from a single run and their sampling errors for the estimates are assessed by the inverse of the information matrix.

Generally, the estimates of allele frequency are not much affected by the degrees of linkage and linkage disequilibrium of markers (Table 4), with consistent results from separate and joint analyses. The estimation precision of recombination fraction and linkage disequilibrium can be much increased when two markers are tightly linked or display low nonrandom association between the allelic frequencies of the markers (Table 4). Both accuracy and precision of parameter estimates from a separate analysis are largely reduced when two markers have loose linkage and strong disequilibrium. However, these can be much improved by using a joint analysis of all the five markers based on a Markov model.

DISCUSSION

The originality of the statistical method proposed in this study is a combined use of the current linkage analysis and linkage disequilibrium-based mapping theory to simultaneously estimate genetic map distances and population genetic associations of markers using random samples drawn from a natural population. Linkage analysis looks for coinheritance of different markers or QTL within a chromosomal region, while linkage disequilibrium looks for differences in the frequency of marker alleles between genotypes of a different marker or different categories of a phenotype. The combined analysis not only can overcome the limitations of linkage analysis, as noted in the Introduction, but also can increase the effectiveness and efficiency of linkage disequilibrium mapping aimed at precise estimation of gene location. The new analytical method can be seen as an extension of linkage disequilibrium mapping for human pedigrees with complete family records toward any types of natural populations.

In this article, a mapping model is developed for dioecious plant species. The progeny of random samples collected from a dioecious population form a series of open-pollinated (or half-sib) families each with a common female parent and different male parents. The experimental strategy for including both the sampled plants and their progeny for genome mapping offers a unique opportunity to study the transmission of genes from the parental to progeny generation, which causes the breakdown of linkage through meiotic recombina-

TABLE 4

MLEs (± SE) of the genetic parameters for five biallelic markers calculated from a single run

Marker	Combination pattern	Hypothesized value			MLE by separate analysis			MLE by joint analysis		
		P_r^i	$\theta^{i(i+1)}$	D_{rs}^{ij}	$\hat{P}_r^i$	$\hat{\theta}^{i(i+1)}$	$\hat{D}_{rs}^{ij}$	$\hat{P}_r^i$	$\hat{\theta}^{i(i+1)}$	$\hat{D}_{rs}^{ij}$
M^1		0.40			0.4051 ± 0.0008			0.4001 ± 0.0004		
	1		0.02	0.02		0.0201 ± 0.0004	0.0205 ± 0.0005		0.0204 ± 0.0004	0.0205 ± 0.0006
M^2		0.40			0.4030 ± 0.0010			0.4008 ± 0.0008		
	2		0.02	0.12		0.0206 ± 0.0012	0.1285 ± 0.0074		0.0193 ± 0.0009	0.1205 ± 0.0034
M^3		0.40			0.4021 ± 0.0015			0.4021 ± 0.0009		
	3		0.30	0.02		0.3102 ± 0.0084	0.0242 ± 0.0056		0.3012 ± 0.0014	0.0215 ± 0.0024
M^4		0.40			0.3974 ± 0.0019			0.4006 ± 0.0007		
	4		0.30	0.12		0.3194 ± 0.0109	0.1105 ± 0.0108		0.3046 ± 0.0054	0.1213 ± 0.0077
M^5		0.40			0.4098 ± 0.0032			0.4035 ± 0.0011		

Standard errors (SE) are derived from the Fisher information index. Combination patterns include (1) tight linkage and weak disequilibrium, (2) tight linkage and strong disequilibrium, (3) loose linkage and weak disequilibrium, and (4) loose linkage and strong disequilibrium.

tion and, thus, the dissipation of linkage disequilibrium between two markers. Unlike previous strategies for a linkage disequilibrium analysis (Hill 1974; Weir and Cockerham 1978), the new strategy, therefore, can capture the intrinsic relationship between linkage and linkage disequilibrium. In human genetic mapping, simultaneous estimation of linkage and linkage disequilibrium is based on nuclear family data (Allison 1997). Such a strategy cannot take advantage of analyzing random samples from a natural population in that one can collect sample sizes as large as those considered in the simulation study. In practice, the new strategy can make an immediate application to many plant species in which progeny tests have been established in the field for a number of years (McKeand and Bridgwater 1998).

Given a fixed sample size, our simulation study has focused on the influence of different allocations of the samples between and within families on parameter estimation. When a sample size is adequately large, for instance, as is that used in our example, the precise estimation of genetic parameters, allele frequencies, linkage, and linkage disequilibrium for markers can be obtained, irrespective of few large families or many small families. Such an advantage for the strategy proposed in this article results from two reasons. First, our mapping analysis is established on the foundation of both parental generation and open-pollinated progeny generation. As a random sample, the parental generation contains as much full information about marker allele frequencies, linkage, and linkage disequilibrium as the original population. Unlike full-sib families, open-pollinated families used in our strategy contain full information not only about marker linkage but also about marker population genetic properties due to the contribution of the paternal gametes (pollen) from the population. Second, our linkage analysis of known marker genotypes includes no missing information, a situation not analogous to QTL mapping in which the genotypes at QTL are unknown.

In this study, we implement the Fisher-scoring algorithm to obtain the MLEs of unknown parameters defining the likelihood function of a marker dataset. The Fisher-scoring algorithm is computationally faster and can be more easily derived (Edwards 1984), as compared to the expectation-maximization (EM) algorithm (Dempster *et al.* 1977) or Markov chain Monte Carlo method (MCMC; Hoeschele *et al.* 1997). Its implementation permits a multiple sampling technique to be used more conveniently. Also, this algorithm can provide the estimates for the asymptotic variances of the parameter estimates. For parameter estimates of known markers with joint genotypes in a multinomial distribution, the asymptotic variances estimated from the Fisher information matrix can adequately describe their sampling errors, especially for large samples, although this may not be an actual case for QTL mapping as seen in Kao and Zeng (1997). In some situations, the calculation of the sampling errors from the Fisher information matrix may encounter negative definitive or singular problems of the matrix. Although these may not be serious for the linkage mapping of known markers, it is advisable to try several different starting values for the parameters to be estimated.

In our experience, a simple Fisher-scoring algorithm is sufficient for analyzing informative markers of known genotypes. However, for a real dataset, there may be many marker types of different segregating patterns. Some markers may be dominant and others may be incomplete or misscored. Many questions for treating these noninformative markers are still open. For example, can we extract useful information from these markers to globally enhance our joint linkage and linkage disequilibrium analysis throughout an entire genome? If yes, how do we make this more efficient? Because of the involvement of the markers of missing information, the Fisher-scoring algorithm may be insufficient for parameter estimation. The EM algorithm or MCMC methodology should be developed to effectively handle these missing data. In addition, when the idea for a combined linkage and linkage disequilibrium analysis is extended to map QTL of unknown genotypes, which is viewed as a missing data problem, the Fisher-scoring algorithm may be very limited. For QTL mapping, more advanced approaches, such as EM algorithm or MCMC, should be developed. Although these approaches are computationally demanding, they can take account of the distribution of multilocus marker-QTL genotypes and permit investigators to fit different models of variation at the QTL.

One of the major contributions of this study is to derive general formulas for estimating allelic frequencies, recombination fractions, and linkage disequilibria for multiallelic markers in natural populations. A number of molecular experiments have demonstrated that multiple alleles per genetic locus are very common in undomesticated populations, such as forest trees (Degen *et al.* 1999). Also, the capacity to detect multiallelic markers is largely enhanced by the development of new biotechnologies such as microsatellites. Analyses of multiallelic markers can be simplified by collapsing them into a few alleles at each locus. But, as found by Weir and Cockerham (1978), this simplification may change the power of detecting linkage disequilibrium and lose some important information about disequilibrium inferences. Thus, it is especially not advisable to use a collapsed set of data when the aim of a study is precise localization of QTL and its subsequent positional cloning.

Our mapping approach here is based on a two-point analysis. We further extend the simple two-point analysis to include all markers from the same chromosome through a Markov model. Such a joint two-point analysis can increase both accuracy and precision of parameter estimation, as demonstrated by a simulation study. For

two markers that are not strongly linked but strongly associated between their allelic frequencies, the two-point analysis excluding other marker information likely has low precision (Table 4). But when other markers are included, the precision of the analysis of these two markers is much increased. Apart from the improvement of the precision of parameter estimation, a joint two-point analysis based on a Markov model can facilitate the ordering of molecular markers on a chromosome (see also LANDER and GREEN 1987). The other means of ordering markers is to develop a multipoint analysis. Although this is mathematically very complicated, the extension of our method to a multipoint analysis is straightforward. Assume that there are three different markers with an order $\mathbf{M}^i$, $\mathbf{M}^j$, and $\mathbf{M}^k$ on the same chromosomes. A three-locus gamete $M^i_r M^j_s M^k_t$ ($r = 1, \ldots, n_i$, $s = 1, \ldots, n_j$, $t = 1, \ldots, n_k$) has the frequency in the original population,

$$P^{ijk}_{rst} = P^i_r P^j_s P^k_t + P^i_r D^{jk}_{st} + P^j_s D^{ik}_{rt} + P^k_t D^{ij}_{rs} + D^{ijk}_{rst},$$

where a three-locus linkage disequilibrium D^{ijk}_{rst} is assumed to exist (WEIR 1996). In a three-point analysis, the relationship between linkage disequilibrium and recombination fraction becomes nonlinear when genes are transmitted from parental to progeny generation (R. L. WU, unpublished results). The gamete frequencies of three-locus gametes in the current and next generations can be expanded to formulate a likelihood function, given the data as in (4).

Many of our species are still in wild states and are of great importance in terms of their economical significance and theoretical values of biological research. For example, as evidenced in TANKSLEY and MCCOUCH (1997), a number of favorable disease-resistant genes in crop plants are currently warehoused in natural populations of their wild relative species and can be made useful to humans if the number, effects, and locations of these genes are understood. From an evolutionary perspective, knowledge about the organization and structure of wild populations helps us to understand the genetic mechanisms of population evolution and make reasonable predictions about the dynamic changes of the populations (BARTON 2000). Unfortunately, the gene-level studies of genetic architecture and inheritance mode for a complex trait in natural populations are surprisingly rare as a result of the paucity of powerful tools to effectively analyze these populations. Although the method reported here is developed for dioecious species, its extension to monoecious species is possible, but requires an additional mathematical manipulation on outcrossing rate, a population genetic parameter that describes the relative importance of outcrossing pollination to selfing pollination within the same plant. With such powerful mapping approaches available to different kinds of species, we are close to addressing many theoretical or practical genetic questions in depth for natural populations.

We are grateful to Dr. George Casella, Dr. Bruce Weir, and Dr. Mark Yang for stimulating discussions about this work; Dr. James Hobert and Dr. Kenneth Portier for helpful readings of this manuscript; and Dr. Gary Churchill and two anonymous referees for thoughtful comments on this manuscript. This research is partially supported by a grant (GM 45344) from the National Institutes of Health.

LITERATURE CITED

ALLISON, D. B., 1997 Transmission disequilibrium tests for quantitative traits. Am. J. Hum. Genet. **60:** 676–690.

BARTON, N. H., 2000 Estimating multilocus linkage disequilibria. Heredity **84:** 373–389.

BROWN, A. D. H., 1975 Sample sizes required to detect linkage disequilibrium between two or three loci. Theor. Popul. Biol. **8:** 184–201.

CAMP, N. J., 1998 Genomewide transmission/disequilibrium testing—consideration of the genotype relative risks at disease loci. Am. J. Hum. Genet. **61:** 1424–1430.

COLLINS, A., and N. E. MORTON, 1998 Mapping a disease locus by allelic association. Proc. Natl. Acad. Sci. USA **95:** 1741–1745.

DARVASI, A., A. WEINREB, V. MINKE, J. I. WELLER and M. SOLLER, 1993 Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics **134:** 943–951.

DEGEN, B., R. STREIFF and B. ZIEGENHAGEN, 1999 Comparative study of genetic variation and differentiation of two pedunculate oak (*Quercus robur*) stands using microsatellite and allozyme loci. Heredity **83:** 597–603.

DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN, 1977 Maximum likelihood from incomplete data via EM algorithm. J. R. Stat. Soc. Ser. B **39:** 1–38.

EDWARDS, A. W. F., 1984 *Likelihood.* Cambridge University Press, Cambridge, UK.

EPPERSON, B. K., and R. W. ALLARD, 1987 Linkage disequilibrium between allozymes in natural populations of lodgepole pine. Genetics **115:** 341–352.

ESCAMILLA, M. A., L. A. MCINNES, M. SPESNY *et al.*, 1999 Assessing the feasibility of linkage disequilibrium methods for mapping complex traits: an initial screen for bipolar disorder loci on chromosome 18. Am. J. Hum. Genet. **64:** 1670–1678.

FARNIR, F., W. COPPIETERS, J-J. ARRANZ, P. BERZI *et al.*, 2000 Extensive genome-wide linkage disequilibrium in cattle. Genome Res. **10:** 220–227.

HÄSTBACKA, J., A. DE LA CHAPELLE, I. KAITILA, P. SISTONEN, A. WEAVER *et al.*, 1992 Linkage disequilibrium mapping in isolated founder populations: diastropic dysplasia in Finland. Nat. Genet. **2:** 204–221.

HÄSTBACKA, J., A. DE LA CHAPELLE, M. MAHTANI, G. CLINES, M. P. REEVE-DALY *et al.*, 1994 The diastropic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell **78:** 1073–1087.

HILL, W. G., 1974 Estimation of linkage disequilibrium in randomly mating populations. Theor. Appl. Genet. **33:** 54–78.

HILL, W. G., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **33:** 54–78.

HOESCHELE, I., P. UIMARI, F. E. GRIGNOLA, Q. ZHANG and K. M. GAGE, 1997 Advances in statistical methods to map quantitative trait loci in outbred populations. Genetics **147:** 1445–1457.

KAO, C-H., and Z-B. ZENG, 1997 General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics **53:** 653–665.

KAPLAN, N. L., W. G. HILL and B. S. WEIR, 1995 Likelihood methods for locating disease genes in nonequilibrium populations. Am. J. Hum. Genet. **56:** 18–32.

KRUGLYAK, L., 1999 Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat. Genet. **22:** 139–144.

LANDEGREN, U., M. NILSSON and P. Y. KWOK, 1998 Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. Genome Res. **8:** 769–776.

LANDER, E. S., and P. GREEN, 1987 Construction of multilocus ge-

netic linkage maps in human. Proc. Natl. Acad. Sci. **84:** 2363–2367.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations for heterotic models. Genetics **49:** 49–67.

Long, A. D., S. L. Mullaney, L. A. Reid, J. D. Fry, C. H. Langley *et al.*, 1995 High resolution mapping of genetic factors affecting abdominal bristle number in *Drosophila melanogaster*. Genetics **139:** 1273–1291.

Luo, Z. W., 1998 Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Heredity **80:** 198–208.

Luo, Z. W., and S. Suhai, 1999 Estimating linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Genetics **151:** 359–371.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Mather, K., and J. L. Jinks, 1982 *Biometrical Genetics*, Ed. 3. Chapman & Hall, New York.

McKeand, S. E., and F. E. Bridgwater, 1998 A strategy for the third breeding cycle of loblolly pine in the Southeastern US. Silvae Genet. **47:** 223–234.

Meuwissen, T. H. E., and M. E. Goddard, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics **155:** 421–430.

Nagylaki, T., 1991 *Introduction to Theoretical Population Genetics*. Springer-Verlag, Berlin.

Nei, M., and W.-H. Li, 1973 Linkage disequilibrium in subdivided populations. Genetics **75:** 213–219.

Ohta, T., 1982a Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc. Natl. Acad. Sci. **79:** 1940–1944.

Ohta, T., 1982b Linkage disequilibrium with the island model. Genetics **101:** 139–155.

Peterson, R. J., D. Goldman and J. C. Long, 1999 Effects of worldwide population subdivision on ALDH2 linkage disequilibrium. Genome Res. **9:** 844–852.

Rabinowitz, D., 1997 A transmission disequilibrium test for quantitative trait loci. Hum. Hered. **47:** 342–350.

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

Service, S. K., D. W. T. Lang, N. B. Freimer and L. A. Sandkuijl, 1999 Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am. J. Hum. Genet. **64:** 1728–1738.

Tanksley, S. D., and S. R. McCouch, 1997 Seed banks and molecular maps: unlocking genetic potential from the wild. Science **277:** 1063–1066.

Templeton, A. R., 1999 Uses of evolutionary theory in the human genome project. Annu. Rev. Ecol. Syst. **30:** 23–49.

Terwilliger, J. D., and K. M. Weiss, 1998 Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr. Opin. Biotechnol. **9:** 578–594.

Wang, D. G., J. B. Fan, C. J. Siao *et al.*, 1998 Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science **280:** 1077–1082.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Weir, B. S., and C. C. Cockerham, 1978 Testing hypotheses about linkage disequilibrium with multiple alleles. Genetics **88:** 633–642.

Xiong, M. M., and S. W. Guo, 1997 Fine-linkage genetic mapping based on linkage disequilibrium: theory and applications. Am. J. Hum. Genet. **60:** 1513–1531.

Communicating editor: G. A. Churchill

APPENDIX A: DERIVATION OF GAMETE FREQUENCIES

When a zygotic genotype is homozygous for both markers $\mathbf{M}^i$ and $\mathbf{M}^j$, only one type of gamete is produced and, thus, recombinant and nonrecombinant gametes are mixed. When a zygotic genotype is homozygous for a marker but heterozygous for the other marker, two types of gametes are produced. In this case, one still cannot distinguish between recombinant and nonrecombinant gametes, because each gamete type is mixed. However, for a genotype that is heterozygous at both markers, four types of gametes can be produced. The frequency for each of the four gamete types produced by a two-marker heterozygous genotype is dependent on the recombination fraction of the markers and the frequency with which the heterozygous genotype was yielded through gamete combination in the previous generation. There are two ways for yielding the heterozygous genotype $M^i_{r_1}M^i_{r_2}M^j_{s_1}M^j_{s_2}$ ($r_1 \neq r_2$, $s_1 \neq s_2$): (1) via the combination of two gametes $M^i_{r_1}M^j_{s_1}$ and $M^i_{r_2}M^j_{s_2}$ and (2) via the combination of two gametes $M^i_{r_1}M^j_{s_2}$ and $M^i_{r_2}M^j_{s_1}$. These two different ways therefore produce two different diplotypes. The probability of the diplotype produced the first way is $2P^{ij}_{r_1s_1}P^{ij}_{r_2s_2}$, whereas the probability of the diplotype produced the second way is $2P^{ij}_{r_1s_2}P^{ij}_{r_2s_1}$ (see Equation 2). The frequencies of the four gamete types produced by the heterozygous genotype are $\frac{1}{2}(1-\theta^{ij})$, $\frac{1}{2}\theta^{ij}$, $\frac{1}{2}\theta^{ij}$, and $\frac{1}{2}(1-\theta^{ij})$ in the first way and $\frac{1}{2}\theta^{ij}$, $\frac{1}{2}(1-\theta^{ij})$, $\frac{1}{2}(1-\theta^{ij})$, and $\frac{1}{2}\theta^{ij}$ in the second way for $M^i_{r_1}M^j_{s_1}$, $M^i_{r_1}M^j_{s_2}$, $M^i_{r_2}M^j_{s_1}$, and $M^i_{r_2}M^j_{s_2}$, respectively. Thus, it is not difficult to derive the frequencies of the four gametes $M^i_{r_1}M^j_{s_1}$, $M^i_{r_1}M^j_{s_2}$, $M^i_{r_2}M^j_{s_1}$, and $M^i_{r_2}M^j_{s_2}$ produced by a heterozygous genotype in the entire population as $P^{ij}_{r_1s_1}P^{ij}_{r_2s_2} - \theta^{ij}D^{ij}$, $P^{ij}_{r_1s_2}P^{ij}_{r_2s_1} + \theta^{ij}D^{ij}$, $P^{ij}_{r_1s_2}P^{ij}_{r_2s_1} + \theta^{ij}D^{ij}$, and $P^{ij}_{r_1s_1}P^{ij}_{r_2s_2} - \theta^{ij}D^{ij}$ (see Table 1), respectively, where $D^{ij} = P^{ij}_{r_1s_1}P^{ij}_{r_2s_2} - P^{ij}_{r_1s_2}P^{ij}_{r_2s_1}$. The conditional probabilities of these gametes are derived according to Bayes' theorem (see Table 1).

APPENDIX B: FISHER-SCORING ALGORITHMS FOR OBTAINING MLES OF $\boldsymbol{\pi}$

For the Fisher-scoring algorithm based on iterative steps, the estimates at the $(\tau+1)$th iteration can be expressed by

$$\boldsymbol{\pi}^{ij[\tau+1]} = \boldsymbol{\pi}^{ij[\tau]} + \mathbf{I}^{-1}(\boldsymbol{\pi}^{ij[\tau]})\mathbf{S}(\boldsymbol{\pi}^{ij[\tau]}),$$

where

$$\mathbf{S}(\boldsymbol{\pi}^{ij[\tau]}) = \frac{\partial}{\partial \boldsymbol{\pi}^{ij}} \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij}) = (S_{P^i_r} S_{P^j_s} S_{\theta^{ij}} S_{D^{ij}_{rs}})^T$$

$$= \left(\frac{\partial \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial P^i_r} \frac{\partial \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial P^j_s} \frac{\partial \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial \theta^{ij}} \frac{\partial \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial D^{ij}_{rs}}\right)^T$$

is the score function, and

$$\mathbf{I}(\boldsymbol{\pi}^{ij}) = -E[\mathbf{S}(\boldsymbol{\pi}^{ij})\mathbf{S}(\boldsymbol{\pi}^{ij})]$$

$$= -\begin{bmatrix} E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial (P^i_r)^2}\right) & E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial P^i_r \partial P^j_s}\right) \\ E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial P^j_s \partial P^i_r}\right) & E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial (P^j_s)^2}\right) \\ E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial \theta^{ij} \partial P^i_r}\right) & E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial \theta^{ij} \partial P^j_s}\right) \\ E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial D^{ij}_{rs} \partial P^i_r}\right) & E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial D^{ij}_{rs} \partial P^j_s}\right) \end{bmatrix}$$

$$\begin{bmatrix} E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial P_r^i \partial \theta^{ij}}\right) & E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial P_r^i \partial D_{rs}^{ij}}\right) \\ E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial P_s^j \partial \theta^{ij}}\right) & E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial P_s^j \partial D_{rs}^{ij}}\right) \\ E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial (\theta^{ij})^2}\right) & E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial \theta^{ij} \partial D_{rs}^{ij}}\right) \\ E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial D_{rs}^{ij} \partial \theta^{ij}}\right) & E\left(\frac{\partial^2 \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})}{\partial (D_{rs}^{ij})^2}\right) \end{bmatrix}$$

is the Fisher information matrix. More specifically, the score function and the Fisher information index can be derived using

$$\mathbf{S}(\boldsymbol{\pi}^{ij\lceil\tau\rceil}) = \frac{\partial}{\partial \boldsymbol{\pi}^{ij}} \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})$$

$$= \prod_{r_1=1}^{n_i} \prod_{r_2=1}^{n_i} \prod_{s_1=1}^{n_j} \prod_{s_2=1}^{n_j} H_{r_1 r_2 s_1 s_2}^{ij} \frac{\partial P_{r_1 r_2 s_1 s_2}^{ij}}{\partial \boldsymbol{\pi}^{ij}} \times \prod_{R_1=1}^{n_i} \prod_{R_2=1}^{n_i} \prod_{S_1=1}^{n_j} \prod_{S_2=1}^{n_j} H_{r_1 r_2 s_1 s_2}^{R_1 R_2 S_1 S_2} \frac{\partial P_{r_1 r_2 s_1 s_2}^{R_1 R_2 S_1 S_2}}{\partial \boldsymbol{\pi}^{ij}},$$

$$\mathbf{I}(\boldsymbol{\pi}^{ij}) = -E\left[\frac{\partial^2}{\partial (\boldsymbol{\pi}^{ij})^2} \ln L(\mathbf{M}|\boldsymbol{\pi}^{ij})\right]$$

$$= -\prod_{r_1=1}^{n_i} \prod_{r_2=1}^{n_i} \prod_{s_1=1}^{n_j} \prod_{s_2=1}^{n_j} H_{r_1 r_2 s_1 s_2}^{ij} \frac{1}{P_{r_1 r_2 s_1 s_2}^{ij}} \left[\frac{\partial P_{r_1 r_2 s_1 s_2}^{ij}}{\partial \boldsymbol{\pi}^{ij}}\right]^2 \times \prod_{R_1=1}^{n_i} \prod_{R_2=1}^{n_i} \prod_{S_1=1}^{n_j} \prod_{S_2=1}^{n_j} H_{r_1 r_2 s_1 s_2}^{R_1 R_2 S_1 S_2} \frac{1}{Q_{r_1 r_2 s_1 s_2}^{R_1 R_2 S_1 S_2}} \left[\frac{\partial Q_{r_1 r_2 s_1 s_2}^{R_1 R_2 S_1 S_2}}{\partial \boldsymbol{\pi}^{ij}}\right]^2.$$