DOI: 10.1534/genetics.103.023044

Simultaneous Detection of Linkage Disequilibrium and Genetic Differentiation of Subdivided Populations

Shuichi Kitada*,[1] and Hirohisa Kishino†

**Faculty of Marine Science, Tokyo University of Marine Science and Technology, Minato, Tokyo 108-8477, Japan and †Graduate School of Agricultural and Life Sciences, University of Tokyo, Bunkyo, Tokyo 113-8657, Japan*

Manuscript received March 20, 2003
Accepted for publication December 26, 2003

ABSTRACT

We propose a new method for simultaneously detecting linkage disequilibrium and genetic structure in subdivided populations. Taking subpopulation structure into account with a hierarchical model, we estimate the magnitude of genetic differentiation and linkage disequilibrium in a metapopulation on the basis of geographical samples, rather than decompose a population into a finite number of random-mating subpopulations. We assume that Hardy-Weinberg equilibrium is satisfied in each locality, but do not assume independence between marker loci. Linkage states remain unknown. Genetic differentiation and linkage disequilibrium are expressed as hyperparameters describing the prior distribution of genotypes or haplotypes. We estimate related parameters by maximizing marginal-likelihood functions and detect linkage equilibrium or disequilibrium by the Akaike information criterion. Our empirical Bayesian model analyzes genotype and haplotype frequencies regardless of haploid or diploid data, so it can be applied to most commonly used genetic markers. The performance of our procedure is examined via numerical simulations in comparison with classical procedures. Finally, we analyze isozyme data of ayu, a severely exploited fish species, and single-nucleotide polymorphisms in human *ALDH2*.

WITH the many discoveries of fine-scale markers that represent highly polymorphic loci, linkage disequilibrium between these markers and disease genes or other trait genes has regained importance (Jorde 1995). Along with the rapid progress of genetic techniques, assessing linkage disequilibrium has become a current concern. The assessment involves two aspects: detecting the presence of disequilibrium and estimating its magnitude once disequilibrium has been confirmed (Weir 1979). Continuous efforts have been made on such assessments of linkage disequilibrium between alleles at two or more loci (Hill 1974a,b; Brown 1975; Weir and Cockerham 1978; Slatkin and Excoffier 1996; Luo 1998; Luo and Suhai 1999; Luo *et al.* 2000; Ayres and Balding 2001; Luo and Wu 2001). Methods for linkage disequilibrium-based mapping of target genes have also been developed (Hill and Weir 1994; Kaplan *et al.* 1995; Xiong and Guo 1997; Meuwissen and Goddard 2000; Wu and Zeng 2001). These methods assume homogeneous natural populations. However, if a population is structured, this leads to biased results (known as spurious association) that can reject a null association between a phenotype and molecular markers (*e.g.*, Lander and Schork 1994).

To overcome this problem, Spielman *et al.* (1993) developed a method called the transmission/disequilibrium test (TDT), which uses nuclear family data to detect real associations in structured populations. To use this type of reliable information, similar family-based methods for testing linkage disequilibrium have been developed (*e.g.*, Excoffier and Slatkin 1998; Lazzeroni and Lange 1998; Spielman and Ewens 1998). However, family-based methods cannot be used for association studies of undomesticated species, such as wildlife and forest trees, for which no nuclear family records are available (Wu and Zeng 2001). Hence, population-based methods that detect linkage disequilibrium even in the presence of population structure are required. Pritchard *et al.* (2000a,b) proposed a population-based method that can detect associations between marker alleles and phenotypes in structured populations. The essential idea of the method is to decompose a sample drawn from a mixed population into several unstructured subpopulations and test the association in the homogeneous subpopulations. The methods have been applied to association analyses in humans (Parra *et al.* 1998; Rosenberg *et al.* 2002) and crop plants, with modified test statistics being used to deal with quantitative traits (Thornsberry *et al.* 2001).

To study the population structure, a sample is often divided by geographical regions. However, in many cases, there are no obvious regional units by which subpopulations can be defined. Rather, natural populations have a complex hierarchical structure, forming a metapopulation (Pannell and Charlesworth 2000). In models,

[1]*Corresponding author:* Tokyo University of Marine Science and Technology, 4-5-7 Konan, Minato, Tokyo 108-8477, Japan.
E-mail: kitada@s.kaiyodai.ac.jp

a metapopulation has a continuous structure and consists of an infinite number of subpopulations, or demes. Such populations are well described by hierarchical models that specify the distribution of genetic structure among demes. Population subdivision (Nei and Li 1973; Peterson *et al.* 1999) and other evolutionary forces such as genetic drift (Hill and Robertson 1968), natural selection (Lewontin 1964), and mutation (Ohta 1982a,b) affect linkage disequilibrium. Therefore, when assessing linkage disequilibrium in natural populations, it is very important to also estimate the magnitude of genetic differentiation.

In this article, we propose a new strategy to detect linkage disequilibrium between marker loci and simultaneously estimate genetic differentiation in metapopulations, extending the empirical Bayesian method of Kitada *et al.* (2000). Instead of decomposing a population into a finite number of subpopulations, we describe a distribution for allele frequencies within subpopulations and test linkage disequilibrium using an information criterion based on hierarchical models that detect either linkage equilibrium or disequilibrium. Using samples from randomly sampled localities from a metapopulation, we estimate hyperparameters associated with linkage disequilibrium and genetic differentiation, on the basis of marginal-likelihood functions derived from prior distributions of allele or haplotype frequencies and likelihood functions. We assume that individuals mate randomly and that Hardy-Weinberg (H-W) equilibrium holds in each locality or deme. The method can be applied to frequency data of common genetic markers, including isozymes, mtDNA, microsatellites, and single-nucleotide polymorphisms (SNPs). We do not assume independence between marker loci, and linkage states remain unknown. We examine the performance of our procedure via numerical simulations in comparison with classical procedures. Finally, we analyze isozyme data of ayu, a severely exploited fish species, and single-nucleotide polymorphisms in human *ALDH2*.

MODELS AND METHODS

Distribution of allele frequencies and genetic differentiation: We consider a metapopulation that consists of localities or demes. Hardy-Weinberg equilibrium holds in each deme. The distribution of the allele frequencies, $\boldsymbol{p} = (p_1, \ldots, p_I)'$, at each deme is described as a Dirichlet distribution (Johnson and Kotz 1969),

$$\pi(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\theta)}{\prod_{i=1}^{I}\Gamma(\alpha_i)}\prod_{i=1}^{I} p_i^{\alpha_i-1}, \tag{1}$$

where I is the number of alleles and $\theta = \sum_{i=1}^{I}\alpha_i$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)'$ are regarded as hyperparameters for respective alleles specifying prior distribution. We use this distribution as a prior for allele frequencies of a subdivided population. According to Weir (1996), for populations that have reached equilibrium under the joint effects of drift and mutation or migration, Wright (1945, 1951) found that allele frequencies for loci with two alleles have a beta distribution; for multiallele loci, the distribution is Dirichlet. We assumed that hyperparameters are common over subpopulations and the sum of the hyperparameters is also common for all loci. So, if random sampling of demes is performed from a metapopulation, estimated hyperparameters describe the magnitude of genetic differentiation among subpopulations that have reached equilibrium. The mean linkage disequilibrium coefficient over subpopulations is also written as a function of the hyperparameters, as shown later. Therefore, our assumption on hyperparameters has genetic meaning.

Now, K demes are randomly sampled from the metapopulation, and n_k individuals are randomly sampled from each deme ($k = 1, \ldots, K$). Given the allele frequencies at each deme, the sample counts of alleles, $\boldsymbol{n_k} = (n_{k1}, \ldots, n_{kI})'$, follow a multinomial distribution. Taking account of uncertainty of the allele frequencies, the likelihood of the sample counts is expressed as a marginal-likelihood function, which is a Dirichlet-multinomial distribution (*e.g.*, Lange 1995; Rannala and Hartigan 1996; Weir 1996; Kitada *et al.* 2000; Balding 2003):

$$\begin{aligned}\tilde{L}(\boldsymbol{\alpha}|\boldsymbol{n_k}) &= \int \ldots \int L(\boldsymbol{p}|\boldsymbol{n_k})\pi(\boldsymbol{p}|\boldsymbol{\alpha})\,d\boldsymbol{p} \\ &= \frac{n_k!}{\prod n_{ki}!}\frac{\Gamma(\theta)}{\Gamma(\theta + n)}\prod_{i=1}^{I}\frac{\Gamma(\alpha_i + n_{ki})}{\Gamma(\alpha_i)}.\end{aligned} \tag{2}$$

The total likelihood is the product of the likelihoods of K demes. Here, $n_k = \sum_{i=1}^{I} n_{ki}$.

With a sample size of n, the variance-covariance matrix of a Dirichlet-multinomial distribution is $(n + \theta)/(1 + \theta)$ times larger than that of the multinomial distribution (Johnson and Kotz 1969). This phenomenon, in which the variance exceeds the nominal variance, is called overdispersion. In our hierarchical models, overdispersion corresponds to the variation of allele frequencies over subpopulations. With samples of size $n_1, \ldots, n_k$, our overdispersion, σ^2, becomes

$$\sigma^2 = \frac{\bar{n} + \theta}{1 + \theta}. \tag{3}$$

Here, $\bar{n}$ is the mean sample count over K samples given by $\bar{n} = \sum_{k=1}^{K} n_k/K$. For a panmictic population, σ^2 is 1; it takes values >1 according to the magnitude of genetic differentiation among subpopulations. On the other hand, θ converges to infinity for a panmictic population.

From Weir's equation (Weir 1996, p. 48, Equation 2.15), the variance of the allele frequency between subpopulations is expressed as

$$\mathrm{Var}[p] = \frac{p(1 - p)}{2n}\{F_{ST}(2n - 1) + 1\}.$$

The term $F_{ST}(2n - 1) + 1$ corresponds to the dispersion parameter σ^2. From this, KITADA *et al.* (2000) obtained the relation between σ^2 and F_{ST} (WRIGHT 1951):

$$F_{ST} = \frac{\sigma^2 - 1}{2n - 1}. \quad (4)$$

If the organism is haploid, $2\overline{n}$ should be $\overline{n}$. By substituting Equation 3 into Equation 4 and assuming $\overline{n} = 2n$, we obtain the relation

$$F_{ST} = \frac{1}{1 + \theta}, \quad (5)$$

which is also given in BALDING (2003). From Equation 5, we have

$$\theta = \frac{1}{F_{ST}} - 1. \quad (6)$$

This coincides with Equation 3 of RANNALA and HARTIGAN (1996), which was proposed by WRIGHT (1969) and gives the rate of gene flow. Thus, the sum of hyperparameters θ is consistent with the rate of gene flow and has a relation with F_{ST}.

Linkage equilibrium with population structure: When genetic data are available from multiple loci, the likelihood is obtained as a product of the likelihoods of these loci when they are under linkage equilibrium. We consider diploid organisms. Let the frequency allele i in a subpopulation at locus j be $p_i^{(j)}$ ($j = 1, \ldots, J$; $i = 1, \ldots, I_j$), where J is the number of loci. The haplotype frequency of the subpopulation under linkage equilibrium can be written by the product of allele frequencies over the loci as

$$h_{i^{(1)}\ldots i^{(J)}} = p_i^{(1)} \ldots p_i^{(J)},$$

where the combination of J alleles $i^{(1)} \ldots i^{(J)}$ should be $i_j^{(1)} \ldots i_j^{(J)}$, but we use ellipses for simplicity. Let the sample count for haplotypes be $n_{i^{(1)}\ldots i^{(J)}}$ and $\Sigma n_{i^{(1)}\ldots i^{(J)}} = 2n$ (individuals). The likelihood for sample haplotypes under H-W equilibrium is a multinomial distribution,

$$L_0(\boldsymbol{p}|\boldsymbol{n}) = C \prod_{i^{(1)}\ldots i^{(J)}} (p_i^{(1)} \cdots p_i^{(J)})^{n_{i^{(1)}\ldots i^{(J)}}} = C \prod_{j=1}^{J} \prod_{i=1}^{I_j} (p_i^{(j)})^{2n_i^{(j)}},$$

where $2n_i^{(j)}$ is the number of genes of allele i at locus j. The constant term $C = 2n!/\Pi_i n_{i^{(1)}\ldots i^{(J)}}!$ is the combination of the observed haplotypes. However, we do not observe haplotypes but genotypes. When composite genotypes $G_{i^{(1)}\ldots i^{(J)}}$ are observed, the likelihood is written as

$$L_0(\boldsymbol{p}|\boldsymbol{n}) = C' 2^{n_{\text{hetero}}} \prod_{i^{(1)}\ldots i^{(J)}} (p_{G_{i^{(1)}\ldots i^{(J)}}})^{n_{G_{i^{(1)}\ldots i^{(J)}}}}, \quad (7)$$

where $C' = n!/\Pi_{i^{(1)}\ldots i^{(J)}} n_{G_{i^{(1)}\ldots i^{(J)}}}!$ and n_{hetero} is the sum of the heterozygous loci over n individuals. Under H-W equilibrium, probabilities for heterozygous genotypes are obtained by duplicating the product of allele frequencies. The term $2^{n_{\text{hetero}}}$ refers to such duplication.

We assume *a priori* that allele frequencies of subpopulations have a Dirichlet (for cases $I_j > 2$) or beta (for cases $I_j = 2$) distribution. The marginal-likelihood function for a subpopulation is obtained from Equations 1, 2, and 7 as

$$\tilde{L}_0(\boldsymbol{\alpha}|\boldsymbol{n}) = C' 2^{n_{\text{hetero}}} \prod_{j=1}^{J} \left\{ \frac{\Gamma(\theta)}{\Gamma(\theta + 2n)} \prod_{i=1}^{I_j} \frac{\Gamma(\alpha_i^{(j)} + 2n_i^{(j)})}{\Gamma(\alpha_i^{(j)})} \right\}, \quad (8)$$

where $2n_i^{(j)}$, the number of genes for allele i at locus j, is calculated from composite genotypes. The total likelihood is the product of K likelihood functions for respec- samples. This equation can also be used for haploid organisms by using n_i and C instead of $2n_i$ and $C'2^{n_{\text{hetero}}}$.

Linkage disequilibrium with population structure: The likelihood for sample haplotypes of a subpopulation is also a multinomial distribution:

$$L_1(\boldsymbol{p}|\boldsymbol{n}) = C \prod_{i^{(1)}\ldots i^{(J)}} (h_{i^{(1)}\ldots i^{(J)}})^{n_{i^{(1)}\ldots i^{(J)}}}. \quad (9)$$

If the haplotypes are observed, the marginal likelihood is given by

$$\tilde{L}_1(\boldsymbol{\alpha}|\boldsymbol{n}) = C \frac{\Gamma(\theta)}{\Gamma(\theta + n)} \prod_{i=1}^{I} \frac{\Gamma(\alpha_i + n_i)}{\Gamma(\alpha_i)}, \quad (10)$$

where I is the number of haplotypes and, for simplicity, we use a suffix i for haplotypes instead of $i^{(1)} \ldots i^{(J)}$. However, when analyzing diploid data, we do not observe diplotypes (and then no haplotypes), but only composite genotypes.

When investigating linkage disequilibrium, the linkage phase between two alleles may be of most interest. Hence, here we focus on models for two loci with two alleles and derive the general marginal-likelihood function in the APPENDIX.

Let allele frequencies for two loci be p_{A_0}, p_{A_1} and p_{B_0}, p_{B_1}, and let the haplotype frequencies be h_{00}, h_{01}, h_{10}, h_{11} in a subpopulation. In this case, nine composite genotypes can be observed with specific diplotypes. Let the number of individuals for the genotypes be $n_1, \ldots, n_9$ (Table 1). For example, the diplotype for the composite genotype $A_0A_0B_1B_1$ can be specified by A_0B_1/A_0B_1. Under the H-W equilibrium assumption, the probability of having the genotype is h_{01}^2 and that for A_0B_0/A_0B_1 is $2h_{00}h_{01}$. Thus, the probabilities of having genotypes can be written by diplotype probabilities, which are combinations of haplotype probabilities. However, for double heterozygotes,

TABLE 1

Notation for number of observed composite genotypes

	B_0/B_0	B_0/B_1	B_1/B_1
A_0/A_0	n_1 (h_{00}^2)	n_2 ($2h_{00}h_{01}$)	n_3 (h_{01}^2)
A_0/A_1	n_4 ($2h_{00}h_{10}$)	n_5 ($2h_{00}h_{11} + 2h_{01}h_{10}$)	n_6 ($2h_{01}h_{11}$)
A_1/A_1	n_7 (h_{10}^2)	n_8 ($2h_{10}h_{11}$)	n_9 (h_{11}^2)

Diplotype probabilities are given in parentheses.

two diplotypes are possible for the genotypes. In this case, $A_0A_1B_0B_1$ is the double heterozygote with observed number n_5. For this genotype, the possible diplotypes are A_0B_0/A_1B_1 and A_0B_1/A_1B_0. The probability for the genotype is then given as $2h_{00}h_{11} + 2h_{01}h_{10}$ (Table 1).

The likelihood of composite genotypes for the case of two loci with two alleles is given in HUDSON (2001, p. 316, Equation 11.6). Expanding the term for the double heterozygote, we obtain the likelihood function as

$$L_1(\boldsymbol{p}|\boldsymbol{n}) = C'2^{n_2+n_4+n_5+n_6+n_8}\sum_{t=0}^{n_5}\binom{n_5}{t}(h_{00})^{2n_1+n_2+n_4+n_5-t}(h_{01})^{n_2+2n_3+n_6+t} \times (h_{10})^{n_4+2n_7+n_8+t}(h_{11})^{n_5+n_6+n_8+2n_9-t}, \quad (11)$$

where $C' = n!/\prod_{i=1}^{9} n_i!$ is a constant term for the combination of the multinomial likelihood. The marginal-likelihood function for a subpopulation is then obtained as

$$\tilde{L}_1(\boldsymbol{\alpha}|\boldsymbol{n}) = C'2^{n_2+n_4+n_5+n_6+n_8}\sum_{t=0}^{n_5}\binom{n_5}{t}\frac{\Gamma(\theta)}{\Gamma(\theta+2n)} \times \frac{\Gamma(\alpha_{00}+2n_1+n_2+n_4+n_5-t)}{\Gamma(\alpha_{00})}\frac{\Gamma(\alpha_{01}+n_2+2n_3+n_6+t)}{\Gamma(\alpha_{01})} \times \frac{\Gamma(\alpha_{10}+n_4+2n_7+n_8+t)}{\Gamma(\alpha_{10})}\frac{\Gamma(\alpha_{11}+n_5+n_6+n_8+2n_9-t)}{\Gamma(\alpha_{11})}, \quad (12)$$

where $\alpha_{11} = \theta - \alpha_{00} - \alpha_{01} - \alpha_{10}$.

Parameter estimation and model selection: We estimate parameters by maximizing the negative log marginal-likelihood functions for linkage equilibrium and disequilibrium. The constant terms C and C' can be excluded from the estimation procedure. We reparameterize hyperparameters with F_{ST} by using the relation given by Equation 6 and estimate F_{ST} and hyperparameters numerically as free parameters by using simplex minimization. The rate of gene flow θ and the dispersion parameter σ^2 are then estimated by Equations 4 and 6. Our empirical Bayesian procedure thus offers maximum-likelihood estimators of genetic differentiation. We estimate the 95% confidence interval for F_{ST} from the log-likelihood profile, and then estimate 95% confidence intervals for θ and σ^2 by substituting the lower and upper confidence limits of F_{ST} into Equations 4 and 6, respectively.

We also estimate the linkage correlation coefficient (HILL and ROBERTSON 1968) as

$$\hat{r} = \frac{\hat{D}_{\text{TOTAL}}}{\sqrt{\hat{E}[p_{A_0}](1-\hat{E}[p_{A_0}])\hat{E}[p_{B_0}](1-\hat{E}[p_{B_0}])}}, \quad (13)$$

where $E[p_{A_0}]$ and $E[p_{B_0}]$ are mean allele frequencies over subpopulations, which are estimated from sample allele frequencies. Here, D_{TOTAL} is the linkage disequilibrium coefficient over subpopulations, $E[h_{00}] - E[p_{A_0}]E[p_{B_0}]$, and is estimated as

$$\hat{D}_{\text{TOTAL}} = \frac{\hat{\alpha}_{00}\hat{\alpha}_{11} - \hat{\alpha}_{01}\hat{\alpha}_{10}}{\hat{\theta}^2}. \quad (14)$$

However, this estimator is biased. Let $E[D]$ be the mean of D at each deme, which can be written as

$$\begin{aligned} E[D] &= E[h_{00} - p_{A_0}p_{B_0}] \\ &= E[h_{00}] - E[p_{A_0}]E[p_{B_0}] + E[p_{A_0}]E[p_{B_0}] - E[p_{A_0}p_{B_0}] \\ &= D_{\text{TOTAL}} - \text{Cov}[p_{A_0}, p_{B_0}], \end{aligned}$$

from which we have

$$D_{\text{TOTAL}} = E[D] + \text{Cov}[p_{A_0}, p_{B_0}]. \quad (15)$$

As $p_{A_0} = h_{00} + h_{01}$ and $p_{B_0} = h_{00} + h_{10}$, the covariance is expressed as

$$\text{Cov}[p_{A_0}, p_{B_0}] = \text{Var}[h_{00}] + \text{Cov}[h_{00}, h_{10}] + \text{Cov}[h_{00}, h_{01}] + \text{Cov}[h_{01}, h_{10}].$$

The variance and covariance of a Dirichlet distribution are $\text{Var}[p_i] = \alpha_i(\theta_i - \alpha_i)/(\theta^2(\theta+1))$ and $\text{Cov}[p_i, p_j] = \alpha_i\alpha_j/(\theta^2(\theta+1))$ (JOHNSON and KOTZ 1969). Hence, we have the covariance as a function of hyperparameters:

$$\widehat{\text{Cov}}[p_{A_0}, p_{B_0}] = \frac{\hat{\alpha}_{00}\hat{\alpha}_{11} - \hat{\alpha}_{01}\hat{\alpha}_{10}}{\hat{\theta}^2(\hat{\theta}+1)}. \quad (16)$$

From Equations 14, 15, and 16, we have the unbiased estimator of the mean linkage disequilibrium coefficient correcting spurious association as

$$\hat{E}[D] = \frac{\hat{\alpha}_{00}\hat{\alpha}_{11} - \hat{\alpha}_{01}\hat{\alpha}_{10}}{\hat{\theta}(\hat{\theta}+1)}, \quad (17)$$

while correction of r is not trivial because of the denominator of Equation 13.

We use the Akaike Information Criterion (AIC; AKAIKE 1973) as a criterion for model selections,

$$\text{AIC} = 2 \times \widehat{LL} + 2 \times u, \quad (18)$$

where $\widehat{LL}$ is the maximum marginal log-likelihood of the model and u is the number of free parameters estimated. We can compare various models by this criterion; the model with the lowest AIC value is selected as the most parsimonious model. Equation 18 indicates that, when there are several models with similar values of the maximum likelihood, we should select the model with the smallest number of parameters, again following the principle of parsimony. By using AIC, we can detect linkage equilibrium or disequilibrium.

TESTING PERFORMANCE

To evaluate the performance of our method, we conducted two sets of simulations. The first simulation examined if our procedure estimates the linkage disequilibrium and genetic differentiation reliably. The second simulation compared our estimate of linkage disequilib-

rium with two other commonly used estimates: an estimate from a pooled sample and an average of the estimated linkage disequilibrium over subsamples.

Simulated data 1: We assumed that the mean population allele frequencies of the two loci over subpopulations were $E[p_{A_0}] = E[p_{B_0}] = 0.5$ and hence haplotype frequencies were $E[h_{00}] = E[h_{11}] = 0.25(1 + r)$ and $E[h_{01}] = E[h_{10}] = 0.25(1 - r)$. We set the sample size to 50 individuals for each sampling point and generated haplotype frequencies for 20 geographical samples under various F_{ST} (0, 0.05, 0.10, 0.15, 0.20) and r (0, 0.2, 0.4, 0.6, 0.8). We then calculated sample composite genotype counts on the basis of haplotype frequencies assuming H-W equilibrium.

For the state of linkage equilibrium ($r = 0$), given the F_{ST} values, sample allele frequencies of two loci were generated independently from the beta distribution $\beta(\theta/2, \theta/2)$, where $\theta = 1/F_{ST} - 1$. Haplotype frequencies were then calculated as $E[p_{A_0}]E[p_{B_0}]$ for $E[h_{00}]$. Other haplotypes were calculated in a similar way. For the case of $F_{ST} = 0$, sample allele frequencies of the two loci were generated independently from the binomial distribution Bi(50, 0.5). For linkage disequilibrium, given the F_{ST} values, haplotype frequencies were generated from the Dirichlet distribution $D(\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11})$, where the hyperparameters were given as $\alpha_{00} = \alpha_{11} = 0.25\theta(1 + r)$ and $\alpha_{01} = \alpha_{10} = 0.25\theta(1 - r)$. For cases of $F_{ST} = 0$, sample haplotype frequencies were generated from the multinomial distribution Mul(50, $E[h_{00}]$, $E[h_{01}]$, $E[h_{10}]$, $E[h_{11}]$).

We estimated F_{ST}, r, and hyperparameters on the basis of Equations 8, 12, and 13 for 1000 replicates. For all cases with $r > 0$, the model of linkage disequilibrium had smaller values of AIC and was selected. On the other hand, for all cases of $r = 0$, the model of linkage equilibrium was selected. Estimates obtained from the best-fit model agreed well with the real values of F_{ST} and r, showing that our method works appropriately over a wide range of genetic differentiation and linkage disequilibrium (Table 2).

Simulated data 2: In an ordinary way, researchers would use the sample mean of the linkage disequilibrium coefficient in each subpopulation (say, D_{MEAN}) as an estimator of the mean linkage disequilibrium coefficient. Although a larger number of sampling points should improve precision of estimates of genetic differentiation, limited sample size in each locality may result in a poor estimate of D_{MEAN}. Another procedure is to estimate the mean linkage disequilibrium coefficient from pooled genotype data over subpopulations (say, D_{POOL}). This procedure neglects the population structure.

The mean population allele frequencies of the two loci over subpopulations were assumed as $E[p_{A_0}] = E[p_{B_0}] = 0.8$ and hence haplotype frequencies were $E[h_{00}] = 0.64 + 0.16r$, $E[h_{01}] = E[h_{10}] = 0.16(1 - r)$, and $E[h_{11}] = 0.04 + 0.16r$. Given F_{ST} values, haplotype frequencies were generated from the Dirichlet distribution, where the hyperparameters were given as $\alpha_{00} = \theta(0.64 + 0.16r)$, $\alpha_{01} = \alpha_{10} = \theta 0.16(1 - r)$, and $\alpha_{11} = \theta(0.04 + 0.16r)$. We set total sample size to 1000 individuals and generated haplotype frequencies for different numbers of localities (K = 10, 20, 40, 100) under various r (0, 0.2, 0.4, 0.6, 0.8) with fixed $F_{ST} = 0.2$. We then calculated sample composite genotype counts on the basis of haplotype frequencies assuming H-W equilibrium.

We estimated F_{ST}, $E[D]$, and hyperparameters on the basis of Equations 8, 12, and 17 for 1000 replicates. D_{MEAN} and D_{POOL} were estimated by the method of Hill (1974b) on the basis of sample composite genotype counts at each locality and the pooled counts. A larger number of sampling points improved the precision of estimates of F_{ST} (Table 3). Our hierarchical model estimated the real values of $E[D]$ correctly over the whole range of linkage disequilibrium and numbers of sampling points. Precision of the estimates of $E[D]$ was also improved for a larger number of sampling points (Table 4). Estimates of D_{MEAN} were biased for weaker linkage disequilibrium. The biases were decreased for larger r; however, the biases were larger than those from the hierarchical model. Estimates of D_{POOL} were largely biased with low precision over the whole range of linkage disequilibrium and numbers of sampling points. The results showed that our method works more efficiently than classical procedures.

APPLICATION TO REAL DATA

Genotype data of the ayu: We analyzed isozyme genotype data for seven samples of the ayu (*Plecoglossus altivelis*) from Japan (K. Yoshizawa, unpublished data). Two samples of wild stocks were taken from Biwako Lake (land-locked type) and Nagara River (amphidoromous type), and five samples were taken from captive brood stocks in hatcheries. The lifespan of ayu is 1 year. These brood stocks have been bred in captivity for between 7 and 26 generations in each hatchery to produce juveniles for release to enhance severely exploited stocks. The numbers of generations of captive breeding are shown in parentheses with the names of the sampling locations in the Table 5 legend. Two loci were analyzed. Four alleles were found at the *Gpi* locus and three alleles at the *Mpi* locus. However, two alleles of *Gpi* and one of *Mpi* were very minor, and so we grouped them with major ones to obtain nine composite genotypes (Table 5).

We estimated parameters on the basis of composite genotype data by using Equations 8, 12, and 13 (Table 6). A smaller AIC value was obtained for the model of linkage equilibrium; thus, linkage equilibrium between *Gpi* and *Mpi* loci was detected. Estimates of F_{ST} and θ showed very large genetic differentiation and small gene flow. Generally, genetic differentiation of fish species is small because of large gene flows caused by migration or lack of barriers (*e.g.*, McPherson *et al.* 2001; Gold and Tur-

TABLE 2

Mean linkage correlation coefficients r and F_{ST} of the best-fit models estimated from 1000 simulations under various levels of linkage disequilibrium and population structure

	Linkage correlation coefficient (r)				
F_{ST}	0.0	0.2	0.4	0.6	0.8
0.00	0 (0[a])	0.200 (0.043)	0.400 (0.039)	0.601 (0.032)	0.802 (0.024)
	0.019 (0.007)	0.019 (0.006)	0.019 (0.006)	0.019 (0.006)	0.020 (0.006)
0.05	0 (0[a])	0.201 (0.056)	0.398 (0.050)	0.600 (0.045)	0.800 (0.032)
	0.047 (0.012)	0.046 (0.010)	0.047 (0.010)	0.047 (0.010)	0.047 (0.011)
0.10	0 (0[a])	0.201 (0.070)	0.398 (0.063)	0.603 (0.053)	0.802 (0.040)
	0.096 (0.021)	0.095 (0.017)	0.094 (0.018)	0.094 (0.018)	0.095 (0.020)
0.15	0 (0[a])	0.203 (0.083)	0.403 (0.075)	0.600 (0.064)	0.803 (0.044)
	0.142 (0.028)	0.142 (0.024)	0.142 (0.024)	0.142 (0.025)	0.142 (0.029)
0.20	0 (0[a])	0.194 (0.095)	0.401 (0.085)	0.603 (0.070)	0.801 (0.048)
	0.190 (0.034)	0.191 (0.028)	0.190 (0.031)	0.192 (0.031)	0.193 (0.037)

The sample size was set to 50 individuals for each sampling point and haplotype frequencies for 20 geographical samples were generated. The mean allele frequencies of the two loci were assumed as $E[p_{A_0}] = E[p_{B_0}] = 0.5$. The numbers in parentheses are the standard errors of the estimates.

[a] In all replicates, the model of linkage equilibrium was selected.

ner 2002). However, the result obtained here is natural because the sample comprised three different groups: land-locked, amphidoromous, and captive brood stocks in successive breeding for many generations.

Human SNPs: We analyzed SNPs reported by Peterson *et al.* (1999) in the human aldehyde dehydrogenase 2 gene *ALDH2*.

We focused on sites 1 and 2 from the six sites reported by Peterson *et al.* The haplotype frequencies were estimated from the haplotype configurations given in their Table 2 as $h_{11} = H_1 + H_4 + H_6 + H_8 + H_9$, $h_{12} = H_3$, $h_{21} = H_2$, and $h_{22} = 0$. We estimated parameters using Equation 8 from the allele frequencies at the two sites within *ALDH2* genotyped in 756 people from 17 populations across five continents (Table 7). We also estimated parameters for estimated haplotypes on the basis of Equation 10 (Tables 7 and 8). The AIC value for the model of linkage disequilibrium was smaller than that for linkage equilibrium. Our estimate of r was -0.60, whereas the original authors' estimates varied from -0.96 to 0.01 for 17 subpopulations, with an average of -0.45 (Peterson *et al.* 1999, Table 4). Strong linkage disequilibrium was found, which is consistent with the analyses by Peterson *et al.* Estimates of F_{ST} and θ also showed that genetic differentiation of human *ALDH2* was large with a small gene flow.

DISCUSSION

We have proposed a new method, based on genotype frequencies of geographical samples, to detect linkage disequilibrium between markers and to simultaneously estimate genetic differentiation by taking population subdivision into account. The method detected linkage disequilibrium and estimated linkage disequilibrium coefficient and F_{ST} correctly.

With a hierarchical model, we estimate the genetic structure in a metapopulation, rather than decompose a population into a finite number of randomly mating subpopulations. The sum of hyperparameters θ coincided

TABLE 3

Mean F_{ST} estimated from 1000 simulations under various levels of linkage disequilibrium for different numbers of sampling points

	$F_{ST} = 0.2$, $E[p_{A_0}] = 0.8$, $E[p_{B_0}] = 0.8$			
r	$K = 10$	$K = 20$	$K = 40$	$K = 100$
0	0.180 (0.047)	0.193 (0.036)	0.195 (0.027)	0.198 (0.020)
0.2	0.181 (0.048)	0.193 (0.037)	0.196 (0.028)	0.198 (0.021)
0.4	0.180 (0.049)	0.190 (0.036)	0.195 (0.028)	0.198 (0.022)
0.6	0.183 (0.052)	0.192 (0.039)	0.197 (0.030)	0.198 (0.023)
0.8	0.183 (0.056)	0.191 (0.044)	0.197 (0.034)	0.198 (0.025)

Total sample size was fixed at 1000 individuals. The numbers in parentheses are the standard errors of the estimates.

TABLE 4

Performance of the hierarchical model in comparison with classical procedures

	$F_{ST} = 0.2$, $E[p_{A_0}] = 0.8$, $E[p_{B_0}] = 0.8$			
$E[D]$ (r)	$K = 10$	$K = 20$	$K = 40$	$K = 100$
0.0	−0.001 (0.014)	0.000 (0.010)	0.000 (0.008)	0.000 (0.006)
(0.0)	−0.000 (0.017)	0.001 (0.012)	0.001 (0.009)	0.003 (0.006)
	−0.001 (0.025)	0.000 (0.018)	0.000 (0.014)	0.000 (0.010)
0.026	0.025 (0.016)	0.025 (0.012)	0.025 (0.009)	0.026 (0.007)
(0.2)	0.042 (0.016)	0.043 (0.075)	0.042 (0.008)	0.042 (0.006)
	0.036 (0.026)	0.033 (0.020)	0.033 (0.015)	0.034 (0.011)
0.051	0.051 (0.014)	0.051 (0.014)	0.052 (0.011)	0.051 (0.007)
(0.4)	0.059 (0.014)	0.059 (0.014)	0.058 (0.010)	0.056 (0.006)
	0.066 (0.022)	0.066 (0.022)	0.067 (0.017)	0.066 (0.011)
0.077	0.077 (0.021)	0.076 (0.016)	0.077 (0.011)	0.077 (0.008)
(0.6)	0.080 (0.022)	0.078 (0.017)	0.077 (0.012)	0.073 (0.007)
	0.096 (0.032)	0.096 (0.024)	0.098 (0.017)	0.097 (0.011)
0.102	0.102 (0.023)	0.102 (0.016)	0.102 (0.012)	0.102 (0.009)
(0.8)	0.102 (0.025)	0.101 (0.018)	0.100 (0.012)	0.094 (0.008)
	0.125 (0.033)	0.127 (0.024)	0.129 (0.016)	0.128 (0.012)

Total sample size was fixed at 1000 individuals. The numbers are means estimated from 1000 simulations with standard errors in parentheses. Lines are arranged in threes as follows: top ($E[D]$), mean of the linkage correlation coefficient at each deme estimated by the hierarchical model; middle (D_{MEAN}), sample mean of the linkage disequilibrium coefficient in each deme; bottom (D_{POOL}), estimates of the linkage disequilibrium coefficient based on pooled genotypes over demes.

with the rate of gene flow, and F_{ST} was written by θ. We also showed by Equation 17 that the magnitude of linkage disequilibrium depends on that of genetic differentiation. Our model assumes metapopulations, which includes Wright's island model as a special case (Wright 1940); hence the relationship between θ and F_{ST} can be applied for these population models. The metapopulation concept fits with ecology of wildlife that have limited movement ability and may mate randomly in their territory. Even for marine fish, which can disperse widely because of the lack of barriers in oceans, the stock concept is popular in fishery resource management (Waples 1998). In breeding seasons, fish species generally gather to spawning grounds where they mate randomly. As a result of such breeding patterns, the assumption of Hardy-Weinberg equilibrium may be valid in each locality. Furthermore, in humans, whose population structures may be more complex than those of wildlife, it has

TABLE 5

Allozyme composite genotype frequencies for seven samples of the ayu from Japan

	Sampling location						
Genotypes	GM-1[a]	GM-2[a]	NG[a]	GF[b]	SG[c]	HS[a]	OT[a]
$A_0A_0B_0B_0$	1	1	65	32	4	3	79
$A_0A_0B_0B_1$	0	2	25	8	7	2	0
$A_0A_0B_1B_1$	0	0	1	2	3	0	0
$A_0A_1B_0B_0$	9	9	7	41	14	30	30
$A_0A_1B_0B_1$	17	18	2	5	27	23	0
$A_0A_1B_1B_1$	9	7	0	2	16	5	0
$A_1A_1B_0B_0$	18	17	0	7	5	19	2
$A_1A_1B_0B_1$	35	36	0	3	18	16	0
$A_1A_1B_1B_1$	11	10	0	1	6	2	0
Total	100	100	100	100	100	100	111

GM, Gunma (26); NG, Niigata (7); GF, Gifu; SG, Shiga; HS, Hiroshima (21); and OT, Oita (11) Prefecture. Numbers in parentheses are the numbers of generations in captive breeding.

[a] Captive brood stocks.

[b] Wild (amphidoromous type).

[c] Wild (land-locked type).

TABLE 6

Estimated linkage correlation coefficient between *Gpi* and *Mpi* loci and genetic differentiation among seven samples of ayu

	Models	
	Linkage equilibrium	Linkage disequilibrium
LL[a]	−1073.25	−1093.89
AIC	2152.50	2195.78
r	—	0.024
F_{ST}	0.280 [0.168, 0.455]	0.209
θ^b	2.576 [1.198, 4.952]	3.787
σ^{2c}	29.164 [17.920, 46.825]	22.039

[a] Maximum log likelihood without the multinomial constant term C' in Equations 8 and 12.
[b] Rate of gene flow.
[c] Dispersion parameter.

TABLE 8

Estimated worldwide linkage correlation coefficient between sites 1 and 2 in the human *ALDH2* gene and genetic differentiation among 17 human populations

	Models	
	Linkage equilibrium	Linkage disequilibrium
LL[a]	−1444.79	−1247.85
AIC	2895.58	2503.70
r	—	−0.601
F_{ST}	0.204	0.234 [0.166, 0.295]
θ	3.908	3.267 [2.390, 5.024]
σ^2	18.917	21.608 [15.598, 26.943]

[a] Maximum log-likelihood without the multinomial constant term C in Equations 8 and 10.

been reported that geographic clusters often correspond closely to predefined regional or population groups or collections of geographically similar populations (Rosenberg *et al.* 2002). In actual populations, various levels of subdivision may exist (Excoffier 2001), and the number of demes should be very large with continuous subdivision. In ecological studies, sampling points may increase year by year, which increases the accuracy of our method.

Estimates of hyperparameters were based only on frequencies counted in the samples, and alleles that did not appear were assigned frequency 0. Here, we consider if this treatment was appropriate. Consider a simple case of three haplotypes with sample counts n_1, n_2, and n_3, where $n_3 = 0$. The marginal-likelihood function for a subpopulation is a Dirichlet-multinomial distribution:

$$L_1(\alpha_1, \alpha_2, \alpha_3 | n_1, n_2, n_3) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3 + n_1 + n_2)} \frac{\Gamma(\alpha_1 + n_1)}{\Gamma(\alpha_1)} \frac{\Gamma(\alpha_2 + n_2)}{\Gamma(\alpha_2)}. \quad (19)$$

Using the relation of $\Gamma(n + 1) = n\Gamma(n)$, the first term of this equation can be written as

$$\frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)(\alpha_1 + \alpha_2 + \alpha_3 + 1) \ldots (\alpha_1 + \alpha_2 + \alpha_3 + n_1 + n_2 - 1)} = \{(\alpha_1 + \alpha_2 + \alpha_3 + 1) \ldots (\alpha_1 + \alpha_2 + \alpha_3 + n_1 + n_2 - 1)\}^{-1}. \quad (20)$$

TABLE 7

Haplotype frequencies at two sites (sites 1 and 2) in the human *ALDH2* gene among 17 human populations, calculated from Table 2 of Peterson *et al.* (1999)

	Alleles		Haplotypes				
Populations	Site 1	Site 2	h_{00}	h_{01}	h_{10}	h_{11}	$2n$
Biaka	10	23	69	23	10	0	102
Cambodian	7	9	33	8	7	0	48
Chinese	10	16	68	16	10	0	94
Japanese	16	10	72	10	16	0	98
S. Korean	13	14	53	14	13	0	80
Taiwanese	10	20	55	21	10	0	86
Black Thai	14	23	63	23	14	0	100
CEPH	49	14	1	14	49	0	64
Finn	65	16	1	16	65	0	82
Swede	76	13	1	13	76	0	90
Cheyenne	66	9	27	9	66	0	102
Mayan	51	11	38	11	51	0	100
Navajo	75	7	10	7	75	0	92
Pima	50	21	19	21	50	0	90
Karitiana	56	19	23	19	56	0	98
R. Surui	83	1	4	1	83	0	88
Ticuna	66	15	22	23	53	0	98
World	717	241	559	249	704	0	1512

Hence, to maximize Equation 19, the term in the parentheses in Equation 20 should be minimized. As $\alpha_3 \geq 0$, the maximum-likelihood estimate of α_3 is then 0. This can be generalized for all cases.

For diploid data, we have two procedures for inferring linkage disequilibrium. The first procedure (method 1) estimates hyperparameters using Equation 12 or the last equation in the APPENDIX, on the basis of observed composite genotypes. Another way (method 2) is to use Equation 10, on the basis of estimated haplotypes. Several methods for estimating haplotype frequencies have been developed (EXCOFFIER and SLATKIN 1995; LONG *et al.* 1995; SLATKIN and EXCOFFIER 1996). In the analysis of human SNPs, we used estimated haplotypes as if they were observed data. However, the effect of using estimated haplotypes on parameter estimates is unknown. The accuracy of the inference might depend on the precision of haplotype estimates. Method 1 should be better for accurate inference, but deriving explicit marginal-likelihood functions becomes much harder for larger numbers of loci, although the general form of the marginal-likelihood function is derived in the APPENDIX. Method 2 is much easier and more practical. Precise estimation of haplotype frequencies is very important for studies of linkage disequilibrium.

We thank Kazutomo Yoshizawa for allowing us to use his unpublished data and Laurent Excoffier and an anonymous reviewer for their constructive comments on earlier versions of the manuscript. This work was supported by the Japan Society for the Promotion of Science and Japan Science and Technology Agency.

LITERATURE CITED

AKAIKE, H., 1973 Information theory and an extension of the maximum likelihood principle, pp. 267–281 in *The Second International Symposium on Information Theory*, edited by B. N. PETROV and F. CSÁKI. Akadémiai Kiadó, Budapest.

AYRES, K. L., and D. J. BALDING, 2001 Measuring gametic disequilibrium from multilocus data. Genetics **157:** 413–423.

BALDING, D. J., 2003 Likelihood-based inference for genetic correlation coefficient. Theor. Popul. Biol. **63:** 221–230.

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

EXCOFFIER, L., 2001 Analysis of population subdivision, pp. 271–307 in *Handbook of Statistical Genetics*, edited by D. J. BALDING, M. BISHOP and C. CANNINGS. Wiley, Chichester, UK.

EXCOFFIER, L., and M. SLATKIN, 1995 Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. **12:** 921–927.

EXCOFFIER, L., and M. SLATKIN, 1998 Incorporating genotypes of relatives into a test of linkage disequilibrium. Am. J. Hum. Genet. **62:** 171–180.

GOLD, J. R., and T. F. TURNER, 2002 Population structure of red drum (*Sciaenops ocellatus*) in the northern Gulf of Mexico, as inferred from variation in nuclear-encoded microsatellites. Marine Biol. **140:** 249–265.

HILL, W. G., 1974a Tests for association of gene frequencies at several loci in random diploid populations. Biometrics **31:** 881–888.

HILL, W. G., 1974b Estimation of linkage disequilibrium in randomly mating populations. Heredity **33:** 229–239.

HILL, W. G., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226–231.

HILL, W. G., and B. S. WEIR, 1994 Maximum likelihood estimation of gene location by linkage disequilibrium. Am. J. Hum. Genet. **54:** 705–714.

HUDSON, R. R., 2001 Linkage disequilibrium and recombination, pp. 309–324 in *Handbook of Statistical Genetics*, edited by D. J. BALDING, M. BISHOP and C. CANNINGS. Wiley, Chichester, UK.

JOHNSON, N. L., and S. KOTZ, 1969 *Discrete Distributions*. Wiley, New York.

JORDE, L. B., 1995 Linkage disequilibrium as a gene mapping tool. Am. J. Hum. Genet. **56:** 11–14.

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

KITADA, S., T. HAYASHI and H. KISHINO, 2000 Empirical Bayes procedure for estimating genetic distance between populations and effective population size. Genetics **156:** 2063–2079.

LANDER, E. S., and N. J. SCHORK, 1994 Genetic dissection of complex traits. Science **265:** 2037–2048.

LANGE, K., 1995 Application of the Dirichlet distribution to forensic match probabilities. Genetica **96:** 107–117.

LAZZERONI, L. C., and K. LANGE, 1998 A conditional inference framework for extending the transmission/disequilibrium test. Hum. Hered. **48:** 67–81.

LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 4–67.

LONG, J. C., R. C. WILLIAMS and M. URBANEK, 1995 An E-M algorithm and testing strategy for multiple-locus haplotypes. Am. J. Hum. Genet. **56:** 799–810.

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.

LUO, Z. W., and C.-I WU, 2001 Modeling linkage disequilibrium between a polymorphic marker locus and a locus affecting complex dichotomous traits in natural populations. Genetics **158:** 1785–1800.

LUO, Z. W., S. H. TAO and Z-B. ZENG, 2000 Inferring linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Genetics **156:** 457–467.

MCPHERSON, A. A., R. L. STEPHENSON, P. T. O'REILLY, M. W. JONES and C. T. TAGGART, 2001 Genetic diversity of coastal Northwest Atlantic herring population implications for management. J. Fish Biol. **59:** 356–370.

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.

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. USA **79:** 1940–1944.

OHTA, T., 1982b Linkage disequilibrium with the island model. Genetics **75:** 139–155.

PANNELL, J. R., and B. CHARLESWORTH, 2000 Effects of metapopulation processes on measures of genetic diversity. Philos. Trans. R. Soc. Lond. B **355:** 1851–1864.

PARRA, E. J., A. MARCINI, J. AKEY, J. MAARTINSON, M. A. BATZER *et al.*, 1998 Estimating African American admixture proportions by use of population-specific alleles. Am. J. Hum. Genet. **63:** 1839–1851.

PETERSON, R. J., D. GOLDMAN and J. C. LONG, 1999 Effects of worldwide population subdivision on *ALDH2* linkage disequilibrium. Genet. Res. **9:** 844–852.

PRITCHARD, J. K., M. STEPHENS, N. A. ROSENBERG and P. DONNELLY, 2000a Association mapping in structured populations. Am. J. Hum. Genet. **67:** 170–181.

PRITCHARD, J. K., M. STEPHENS and P. DONNELLY, 2000b Inference of population structure using multilocus genotype data. Genetics **156:** 945–959.

RANNALA, B., and J. A. HARTIGAN, 1996 Estimating gene flow in island populations. Genet. Res. **67:** 147–158.

ROSENBERG, N. A., J. K. PRITCHARD, J. L. WEBER, H. M. CANN, K. K. KIDD *et al.*, 2002 Genetic structure of human populations. Science **298:** 2381–2384.

SLATKIN, M., and L. EXCOFFIER, 1996 Testing for linkage disequilibrium in genotype data using the EM algorithm. Heredity **76:** 377–383.

SPIELMAN, R. S., and W. J. EWENS, 1998 A sibship test for linkage

in the presence of association: the sib transmission/disequilibrium test. Am. J. Hum. Genet. **62:** 450–458.

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

Thornsberry, J. M., M. M. Goldman, J. Doebley, S. Kresovich, D. Nielsen *et al.*, 2001 Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. **28:** 286–289.

Waples, R. S., 1998 Separating the wheat from the chaff. Pattern of genetic differentiation in high gene flow species. J. Hered. **89:** 438–450.

Weir, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35:** 235–254.

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.

Wright, S., 1940 Breeding structure of populations in relation to speciation. Am. Nat. **74:** 232–248.

Wright, S., 1945 The differential equation of the distribution of gene frequencies. Proc. Natl. Acad. Sci. USA **31:** 383–389.

Wright, S., 1951 The general structure of populations. Ann. Eugen. **15:** 323–354.

Wright, S., 1969 *Evolution and Genetics of Populations: The Theory of Gene Frequencies*, Vol. 2. University of Chicago Press, Chicago.

Wu, R., and Z-B. Zeng, 2001 Joint linkage and linkage disequilibrium mapping in natural populations. Genetics **157:** 899–909.

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: L. Excoffier

APPENDIX: GENERAL LINKAGE DISEQUILIBRIUM MARGINAL-LIKELIHOOD FUNCTION FOR DIPLOID ORGANISMS

Let observed composite genotypes be $G_i (i = 1, \ldots, m)$ and the numbers of individuals for the genotypes be n_{G_i}. The likelihood function for a sample is written as

$$L_1 = C' \prod_{i=1}^{m} P_{G_i}^{n_{G_i}},$$

where P_{G_i} is the probability for the genotype G_i, which can be written by the probability of diplotypes. For the case of two loci with two alleles, the diplotype probabilities are given in Table 1. Even for multilocus data, a diplotype can be specified for a set of completely homozygous genotypes, but the larger number of heterozygous genotypes results in a larger number of diplotype combinations. Consequently, the number of summations of diplotype probabilities increases for such genotypes. Thus, the probability of a genotype is expressed by the sum of diplotype probabilities.

Let a set of diplotypes and haplotypes consistent with a genotype G be

$$\begin{aligned} \text{diplo}(G) &= (D_1^{(G)}, \ldots, D_{l_G}^{(G)})' \\ &= (h_{11}^{(G)}/h_{12}^{(G)}, \ldots, h_{l_G 1}^{(G)}/h_{l_G 2}^{(G)})', \end{aligned}$$

where l_G is the number of diplotypes consistent with the genotype G. The likelihood function for composite genotypes is then written by using diplotype probabilities as

$$L_1 = C' \prod_{i=1}^{m} \left(\sum_{j=1}^{l_{G_i}} P_{D_j} \right)^{n_{G_i}}.$$

By multinomial expansion, we have another form of the likelihood as

$$L_1 = C' \sum_{i_1 \ldots i_{l_{G_1}}} \frac{n_{G_1}!}{i_1^{(1)}! \ldots i_{l_{G_1}}^{(1)}!} \left(P_{D_1^{(G_1)}}^{i_1^{(1)}} \ldots P_{D_{l_{G_1}}^{(G_1)}}^{i_{l_{G_1}}^{(1)}} \right) \ldots$$

$$\sum_{i_1 \ldots i_{l_{G_m}}} \frac{n_{G_m}!}{i_1^{(m)}! \ldots i_{l_{G_m}}^{(m)}!} \left(P_{D_1^{(G_m)}}^{i_1^{(m)}} \ldots P_{D_{l_{G_m}}^{(G_m)}}^{i_{l_{G_m}}^{(m)}} \right),$$

where $i_j^{(i)}$ is the number of individuals having diplotype j, which constitutes genotype i, and $d\Sigma_{j=1}^{l_{G_i}} i_j^{(i)} = n_{G_i}$.

Under H-W equilibrium, the probability of the jth diplotype is expressed as a product of the probabilities of haplotypes,

$$P_{h_{j1}/h_{j2}} = 2^{(1-\delta_{h_{j1}h_{j2}})} P_{h_{j1}} P_{h_{j2}},$$

where δ is a delta function that is 1 for homozygous and 0 for heterozygous genotypes. Therefore, the likelihood of a sample from a subpopulation under H-W equilibrium is

$$\begin{aligned} L_1 = C' \sum_{i_1 \ldots i_{l_{G_1}}} \frac{n_{G_1}!}{i_1^{(1)}! \ldots i_{l_{G_1}}^{(1)}!} &\left\{ \left(2^{(1-\delta_{h_{11}^{(G_1)} h_{12}^{(G_1)}})} P_{h_{11}^{(G_1)}} P_{h_{12}^{(G_1)}} \right)^{i_1^{(1)}} \ldots \right. \\ &\left. \left(2^{(1-\delta_{h_{l_{G_1}1}^{(G_1)} h_{l_{G_1}2}^{(G_1)}})} P_{h_{l_{G_1}1}^{(G_1)}} P_{h_{l_{G_1}2}^{(G_1)}} \right)^{i_{l_{G_1}}^{(1)}} \right\} \ldots \sum_{i_1 \ldots i_{l_{G_m}}} \frac{n_{G_m}!}{i_1^{(m)}! \ldots i_{l_{G_m}}^{(m)}!} \\ &\times \left\{ \left(2^{(1-\delta_{h_{11}^{(G_m)} h_{12}^{(G_m)}})} P_{h_{11}^{(G_m)}} P_{h_{12}^{(G_m)}} \right)^{i_1^{(m)}} \ldots \left(2^{(1-\delta_{h_{l_{G_m}1}^{(G_m)} h_{l_{G_m}2}^{(G_m)}})} P_{h_{l_{G_m}1}^{(G_m)}} P_{h_{l_{G_m}2}^{(G_m)}} \right)^{i_{l_{G_m}}^{(m)}} \right\}. \end{aligned}$$

In the above equation some haplotypes will be common, so we rename these haplotypes as

$$\{h'_1, \ldots, h'_\xi\} = \{h_{11}^{(G_1)}, \ldots, h_{l_{G_m}2}^{(G_m)}\},$$

where ξ is the number of common haplotypes. The likelihood function is then written as

$$\begin{aligned} L_1 = C' \sum_{i_1 \ldots i_{l_{G_1}}} \ldots \sum_{i_1 \ldots i_{l_{G_m}}} &\frac{n_{G_1}!}{i_1^{(1)}! \ldots i_{l_{G_1}}^{(1)}!} \ldots \frac{n_{G_m}!}{i_1^{(m)}! \ldots i_{l_{G_m}}^{(m)}!} \\ &\times \left\{ 2^{(1-\delta_{h_{11}^{(G_1)} h_{12}^{(G_1)}})} \right\}^{i_1^{(1)}} \ldots \left\{ 2^{(1-\delta_{h_{l_{G_m}1}^{(G_m)} h_{l_{G_m}2}^{(G_m)}})} \right\}^{i_{l_{G_m}}^{(m)}} (P_{h'_1})^{j_1} \ldots (P_{h'_\xi})^{j_\xi}, \end{aligned}$$

where j_s $(s = 1, \ldots, \xi)$ is the number of individuals having haplotype s and $\Sigma j_s = 2n$,

$$\begin{aligned} j_s = i_1^{(1)} \left(\delta_{h'_s h_{11}^{(G_1)}} + \delta_{h'_s h_{12}^{(G_1)}} \right) + \ldots + i_{l_{G_1}}^{(1)} \left(\delta_{h'_s h_{l_{G_1}1}^{(G_1)}} + \delta_{h'_s h_{l_{G_1}2}^{(G_1)}} \right) + \ldots \\ + i_1^{(m)} \left(\delta_{h'_s h_{11}^{(G_m)}} + \delta_{h'_s h_{12}^{(G_m)}} \right) + \ldots + i_{l_{G_m}}^{(m)} \left(\delta_{h'_s h_{l_{G_m}1}^{(G_m)}} + \delta_{h'_s h_{l_{G_m}2}^{(G_m)}} \right), \end{aligned}$$

where $\delta_{h'_s h_{11}^{(G_1)}}$ is again a delta function that is 1 if h'_s coincides with $h_{11}^{(G_1)}$ and 0 otherwise. Thus, the likelihood for observed composite genotypes can be written as the sum of the multinomial distribution for haplotype frequencies. From above L_1, the marginal-likelihood function is obtained as

$$\tilde{L}_1(\boldsymbol{\alpha}|\boldsymbol{n}) = C' \sum_{i_1 \ldots i_{l_{G_1}}} \cdots \sum_{i_1 \ldots i_{l_{G_m}}} \frac{n_{G_1}!}{i_1^{(1)}! \ldots i_{l_{G_1}}^{(1)}!} \cdots \frac{n_{G_m}!}{i_1^{(m)}! \ldots i_{l_{G_m}}^{(m)}!} \times \left\{2^{(1-\delta_{h_{11}^{(G_1)} h_{12}^{(G_1)}})}\right\}^{i_1^{(1)}} \cdots \left\{2^{(1-\delta_{h_{l_{G_m}1}^{(G_m)} h_{l_{G_m}2}^{(G_m)}})}\right\}^{i_{l_{G_m}}^{(m)}} \times \left\{\frac{\Gamma(\theta)}{\Gamma(\theta + 2n)} \prod_{s=1}^{\xi} \frac{\Gamma(\alpha_s + j_s)}{\Gamma(\alpha_s)}\right\}.$$

The total marginal-likelihood function is obtained multiplying K likelihood functions.