Genetics: Published Articles Ahead of Print, published on June 11, 2007 as 10.1534/genetics.106.068890

Estimation of Multilocus Linkage Disequilibria in Diploid Populations with Dominant Markers

Yanchun Li,* Yang Li,† Kun Han,* Zhengjia Wang,* Wei Hou,‡ Yanru Zeng* Rongling Wu†*

School of Forestry and Biotechnology, Zhejiang Forestry University, Lin'an, Zhejiang 311300 People's Republic of China, ‡Department of Epidemiology and Health Policy Research, †Department of Statistics, University of Florida, Gainesville, FL 32611 USA

Submitted to *Genetics*

Running Head: Linkage Disequilibrium Analysis

Keywords: Linkage disequilibrium, Dominant marker, Codominant marker, Natural population, EM algorithm

Corresponding author:

Rongling Wu

Department of Statistics

University of Florida

Gainesville, FL 32611

Tel: (352)392-3806

Fax: (352)392-8555

Email: rwu@stat.ufl.edu

ABSTRACT

Analysis of population structure and organization with DNA-based markers can provide important information regarding the history and evolution of a species. Linkage disequilibrium (LD) analysis based on allelic associations between different loci is emerging as a viable tool to unravel the genetic basis of population differentiation. In this article, we derive the EM algorithm to obtain the maximum likelihood estimates of the linkage disequilibria between dominant markers, aimed to study the patterns of genetic diversity for a diploid species. The algorithm was expanded to estimate and test linkage disequilibria of different orders among three dominant markers and can be technically extended to manipulate an arbitrary number of dominant markers. The algorithm is validated by a real example for population genetic studies of hickory trees, native to the southeastern China, using dominant markers. Extensive simulation studies were performed to investigate the statistical properties of this algorithm. The precision of our estimates of linkage disequilibrium between dominant markers were compared between that between codominant markers. Results from simulation studies suggest that three-locus LD analysis can increase the power for LD detection as compared with two-locus LD analysis. The algorithm will be useful for studying the pattern and amount of genetic variation within and among populations.

INTRODUCTION

The pattern and extent of non-random associations among polymorphic markers distributed over the genome are related to the evolutionary rate of population structure for a species (Tishkoff et al. 1996, 2001; Stephens et al. 2001; Weiss and Clark 2002; Ardlie et al. 2002). One measure of such non-random associations, linkage disequilibrium (LD), is often affected by various evolutionary forces, such as selection, genetic drift, mutation, admixture, population structure and etc., which have operated in the population. For this reason, by estimating and testing for the extent and distribution of LD throughout the genome, the

evolution of population structure can be inferred (Tishkoff et al. 2002). There has been a wealth of literature on the application of the LD analysis to understand the population evolution of humans (Reich et al. 2001; Dawson et al. 2002; Gabriel et al. 2002; Ardlie et al. 2002) as well as a variety of plants and animals (Remington et al. 2001; Hansson et al. 2004; Liu et al. 2006).

Unlike humans and several model systems, such as mouse and Arabidopsis, in which high-resolution LD maps have been constructed with codominant markers, such as single nucleotide polymorphisms (SNPs) and microsatellites, many underrepresented species, like forest trees, still heavily rely upon simple and cheap dominant marker techniques. These markers including random amplified polymorphic DNA (RAPD, Williams et al. 1990) and amplified fragment length polymorphism (AFLP, Vos et al. 1995) can be genotyped arbitrarily from the genome with no need of prior knowledge about the structure and sequence of the genome. There have been extensive publications on the use of dominant markers to explore the amount, structure, and distribution of genetic variation in a population (Yan et al. 1999; Zhivotovsky 1999; Holsinger et al. 2002; Miller and Schaal 2006) and manage biological resources and diversity in agriculture and forestry (Kuang et al. 1998; Silbiger et al. 1998; Kremer et al. 2005). In the post-genomic era, proteomic techniques have become increasingly available to produce dominant markers such as presence/absence sport (PAS) and protein quantitative locus (PQL) (Thiellement et al. 1999; Zivy and Vienne 2000; Consoli et al. 2002). Proteomic markers are often related to biological functions and, therefore, will play an important role in the genetic study of variation in a natural or experimental population.

For a dominant marker, the plus-allele, shown as the presence of a band on the gel, dominates over the null-allele, shown as the absence of a band (unamplifiable by PCR). Therefore, heterozygous and homozygous genotypes both with the band cannot be directly distinguished from each other. Statistical methods for estimating linkage disequilibria with codominant markers, for which two homozygotes and the heterozygote are distinguishable,

have well been developed (Pritchard and Przeworski 2001; Wang and Wu 2004). However, the estimates of LD among dominant markers have been poorly explored, although this has attracted the interest of geneticists for a long time because of its potential importance in the understanding of population structure (Hill 1974). As compared to codominant markers, dominant markers are less informative and, therefore, there may be some loss of information for their practical use. In this article, we will derive a statistical method for estimating and testing the LD between dominant markers within the maximum likelihood context. This method is implemented with the EM algorithm, allowing for a number of hypotheses regarding the structure and organization of a population. The statistical properties of the method for analyzing dominant markers are investigated and compared with those for codominant markers through simulation studies.

TWO-LOCUS LINKAGE DISEQUILIBRIUM ANALYSIS

The description of our LD analysis model will start with more informative codominant markers. The principle for analyzing codominant markers is then expanded to derive the model for LD analyses of less informative dominant markers.

The Codominant Model: Suppose there are n diploid individuals that are randomly drawn from a large population at Hardy-Weinberg equilibrium. Let us consider two diallelic markers **A** (with alleles A and a) and **B** (with two alleles B and b). These two markers are genetically associated with the linkage disequilibrium D in an HWE population. Allele frequencies are defined as p_A and p_a ($p_A + p_a = 1$) for **A** and p_B and p_b ($p_B + p_b = 1$) for **B**, respectively. The frequencies of four haplotypes, AB, Ab, aB and ab, formed by the two markers, are denoted, respectively, as

$$\begin{array}{rcl} p_{AB} & = & p_A p_B + D \\ p_{Ab} & = & p_A p_b - D \\ p_{aB} & = & p_a p_B - D \\ p_{ab} & = & p_a p_b + D, \end{array} \tag{1}$$

summing to one.

Each of the two markers has three genotypes, that is, AA, Aa and aa for marker **A**, and BB, Bb and bb for marker **B**. Let $n_{l_\mathbf{A} l_\mathbf{B}}$ be the observed number of observations of a two-marker genotype $l_\mathbf{A} l_\mathbf{B}$ ($l_\mathbf{A} = AA, Aa, aa; l_\mathbf{B} = BB, Bb, bb$). For a practical data set, the observations of these genotypes and their expected frequencies can be tabulated in a format like Table 1. Under the HWE, the diplotype frequencies are the products of the corresponding haplotype frequencies. The resulting genotype frequencies for each cell are then expressed in terms of haplotype frequencies.

The likelihood of unknown haplotype frequencies, $\mathbf{\Omega} = (p_{AB}, p_{Ab}, p_{aB}, p_{ab})$, given observed genotypes ($\mathbf{M}$) can be written as a multinomial form, i.e.,

$$\begin{aligned} L(\mathbf{\Omega}|\mathbf{M}) \propto\ & 2n_{AABB}\log p_{AB} + n_{AABb}\log(2p_{AB}p_{Ab}) + 2n_{AAbb}\log p_{Ab} \\ & + n_{AaBB}\log(2p_{AB}p_{Ab}) + n_{AaBb}\log[2(p_{AB}p_{ab} + p_{Ab}p_{aB})] + n_{Aabb}\log(2p_{Ab}p_{ab}) \\ & + 2n_{aaBB}\log p_{aB} + n_{aaBb}\log(2p_{aB}p_{ab}) + 2n_{aabb}\log p_{ab}. \end{aligned} \tag{2}$$

The MLEs of haplotype frequencies are then derived as as

$$\begin{aligned} \widehat{p}_{AB} &= \frac{1}{2n}(2n_{AABB} + n_{AABb} + n_{AaBB} + \phi n_{AaBb}), \\ \widehat{p}_{Ab} &= \frac{1}{2n}[2n_{AAbb} + n_{AABb} + n_{Aabb} + (1-\phi)n_{AaBb}], \\ \widehat{p}_{aB} &= \frac{1}{2n}[2n_{aaBB} + n_{AaBB} + n_{aaBb} + (1-\phi)n_{AaBb}], \\ \widehat{p}_{ab} &= \frac{1}{2n}(2n_{aabb} + n_{Aabb} + n_{aaBb} + \phi n_{AaBb}), \end{aligned} \tag{3}$$

where

$$\phi = \frac{p_{AB}p_{ab}}{p_{AB}p_{ab} + p_{Ab}p_{aB}}. \tag{4}$$

Equations 3 and 4 comprise a loop for the EM algorithm. Initial values for the haplotype frequencies are provided to calculate the proportion ϕ in the E step (4). The calculated ϕ value is then used to estimate haplotype frequencies in the M step (3). Both the E and M

steps are repeated until the estimates of haplotype frequencies converge. The MLEs of allele frequencies at the two markers and their linkage disequilibrium can be obtained by solving a group of equations (1), i.e.,

$$\begin{aligned} \widehat{p}_A &= \widehat{p}_{AB} + \widehat{p}_{Ab} \\ \widehat{p}_B &= \widehat{p}_{AB} + \widehat{p}_{aB} \\ \widehat{D} &= \widehat{p}_{AB}\widehat{p}_{ab} - \widehat{p}_{Ab}\widehat{p}_{aB} \end{aligned} \tag{5}$$

The sampling variances for the MLEs of haplotype frequencies (and therefore allele frequencies and linkage disequilibrium) can be calculated from the Fisher information matrix.

The degree of linkage disequilibrium between two markers can be tested by formulating two hypotheses expressed as

$$H_0 : D = 0 \text{ vs. } D \neq 0, \tag{6}$$

under which the likelihoods, $L(\widetilde{\mathbf{\Omega}}|\mathbf{M})$ and $L(\widehat{\mathbf{\Omega}}|\mathbf{M})$, are calculated, respectively, where the tilde corresponds to the MLEs for the null hypothesis and the hat corresponds to the MLEs for the alternative hypothesis. Under the null hypothesis above, the allele frequencies of codominant markers can be directly estimated from the marker data without the use of the EM algorithm. The log-likelihood ratio test statistic is calculated by

$$\text{LR} = -2[\ln L(\widetilde{\mathbf{\Omega}}|\mathbf{M}) - \ln L(\widehat{\mathbf{\Omega}}|\mathbf{M})] \tag{7}$$

which is asymptotically χ^2-distributed with one degree of freedom.

The Dominant Model: For a dominant marker (say **A**), the homozygote (AA) for the dominant allele cannot be distinguished from the heterozygote (Ao). Thus, these two genotypes are observed as a single "phenotype" ($A\_$). Because of this, some cells for the observations and expected genotype frequencies in Table 1 are collapsed in a way as shown in Table 2. We use $n_{k_\mathbf{A}/k_\mathbf{B}}$ to denote the observed number of observations of a two-dominant-marker genotype $k_\mathbf{A}/k_\mathbf{B}$ [$k_\mathbf{A} = A\_\ (1), oo\ (0); k_\mathbf{B} = B\_\ (1), oo\ (0)$]. Using the same idea as conceived for the codominant markers, we formulate the EM algorithm to estimate haplotype

frequencies

$$\begin{aligned}
\widehat{p}_{AB} &= \frac{1}{2n}(\phi_1 n_{1/1}) \\
\widehat{p}_{Ab} &= \frac{1}{2n}(\phi_2 n_{1/1} + \phi_3 n_{1/0}) \\
\widehat{p}_{aB} &= \frac{1}{2n}(\phi_4 n_{1/1} + \phi_5 n_{0/1}) \\
\widehat{p}_{ab} &= \frac{1}{2n}(\phi_6 n_{1/1} + \phi_7 n_{1/0} + \phi_8 n_{0/1} + 2n_{0/0})
\end{aligned} \tag{8}$$

where

$$\begin{aligned}
\phi_1 &= \frac{2p_{AB}}{2p_{AB} + 2p_{Ab}p_{aB} - p_{AB}^2} \\
\phi_2 &= \frac{2p_{Ab}(p_{AB} + p_{aB})}{2p_{AB} + 2p_{Ab}p_{aB} - p_{AB}^2} \\
\phi_3 &= \frac{2(p_{Ab} + p_{ab})}{p_{Ab} + 2p_{ab}} \\
\phi_4 &= \frac{2p_{aB}(p_{AB} + p_{Ab})}{2p_{AB} + 2p_{Ab}p_{aB} - p_{AB}^2} \\
\phi_5 &= \frac{2(p_{aB} + p_{ab})}{p_{aB} + 2p_{ab}} \\
\phi_6 &= \frac{2p_{AB}p_{ab}}{2p_{AB} + 2p_{Ab}p_{aB} - p_{AB}^2} \\
\phi_7 &= \frac{2p_{ab}}{p_{Ab} + 2p_{ab}} \\
\phi_8 &= \frac{2p_{ab}}{p_{aB} + 2p_{ab}}
\end{aligned} \tag{9}$$

The E (9) and M step (8) are iterated until convergence. The MLEs of the allele frequencies and linkage disequilibrium are obtained by Equation 5. The linkage disequilibrium is similarly tested with hypothesis (6). A similar iterative algorithm for estimating the LD between dominant markers was derived by Hill (1974).

It should be pointed out that, unlike the codominance case, the estimation of allele frequencies of two dominant markers under the null hypothesis of no linkage disequilibrium should be based on the EM algorithm. The EM algorithm is described as follows:

In the E step, calculate

$$\phi_A = \frac{2}{1 + p_a}, \ \phi_a = \frac{2p_a}{1 + p_a},$$

$$\phi_B = \frac{2}{1+p_b},\ \phi_b = \frac{2p_b}{1+p_b}$$

and then in the M step, estimate the allele frequencies of markers **A** and **B** by using

$$p_A = \frac{\phi_A(n_{1/1}+n_{1/0})}{2n},\ p_a = \frac{2(n_{0/1}+n_{0/0})+\phi_a(n_{1/1}+n_{1/0})}{2n},$$
$$p_B = \frac{\phi_B(n_{1/1}+n_{0/1})}{2n},\ p_b = \frac{2(n_{1/0}+n_{0/0})+\phi_b(n_{1/1}+n_{0/1})}{2n}.$$

THREE-LOCUS LINKAGE DISEQUILIBRIUM ANALYSIS

The Codominant Model: Consider three segregating markers, **A** with alleles A and a, **B** with alleles B and b, and **C** with alleles C and c. The notation for marker alleles and their frequencies is given similarly for two markers in the preceding section. Linkage disequilibrium analysis among three markers is characterized by allele frequencies at each marker and linkage disequilibria between any pair of markers and among all the three markers are denoted as $D_{\mathbf{AB}}$, $D_{\mathbf{AC}}$, $D_{\mathbf{BC}}$ and $D_{\mathbf{ABC}}$, respectively (Bennett 1954; Nielsen et al. 2004). Three markers generate eight different haplotypes, expressed as ABC, ABc, AbC, Abc, aBC, aBc, abC and abc, whose frequencies are expressed, respectively, as

$$\begin{array}{lcllllll}
p_{ABC} &=& p_Ap_Bp_C & +p_AD_{\mathbf{BC}} & +p_BD_{\mathbf{AC}} & +p_CD_{\mathbf{AB}} & +D_{\mathbf{ABC}} \\
p_{ABc} &=& p_Ap_Bp_c & -p_AD_{\mathbf{BC}} & -p_BD_{\mathbf{AC}} & +p_cD_{\mathbf{AB}} & -D_{\mathbf{ABC}} \\
p_{AbC} &=& p_Ap_bp_C & -p_AD_{\mathbf{BC}} & +p_bD_{\mathbf{AC}} & -p_CD_{\mathbf{AB}} & -D_{\mathbf{ABC}} \\
p_{Abc} &=& p_Ap_bp_c & +p_AD_{\mathbf{BC}} & -p_bD_{\mathbf{AC}} & -p_cD_{\mathbf{AB}} & +D_{\mathbf{ABC}} \\
p_{aBC} &=& p_ap_Bp_C & +p_aD_{\mathbf{BC}} & -p_BD_{\mathbf{AC}} & -p_CD_{\mathbf{AB}} & -D_{\mathbf{ABC}} \\
p_{aBc} &=& p_ap_Bp_c & -p_aD_{\mathbf{BC}} & +p_BD_{\mathbf{AC}} & -p_cD_{\mathbf{AB}} & +D_{\mathbf{ABC}} \\
p_{abC} &=& p_ap_bp_C & -p_aD_{\mathbf{BC}} & -p_bD_{\mathbf{AC}} & +p_CD_{\mathbf{AB}} & +D_{\mathbf{ABC}} \\
p_{abc} &=& p_ap_bp_c & +p_aD_{\mathbf{BC}} & +p_bD_{\mathbf{AC}} & +p_cD_{\mathbf{AB}} & -D_{\mathbf{ABC}}.
\end{array} \quad (10)$$

These haplotype frequencies are used to describe the diplotype (and genotype) frequencies at the three markers given in Table 3. Let $n_{l_{\mathbf{A}}l_{\mathbf{B}}l_{\mathbf{C}}}$ be the observed number of observations of a three-marker genotype $l_{\mathbf{A}}l_{\mathbf{B}}l_{\mathbf{C}}$ ($l_{\mathbf{A}} = AA, Aa, aa; l_{\mathbf{B}} = BB, Bb, bb; l_{\mathbf{C}} = CC, Cc, cc$). Based on the expected frequencies, generally expressed as $P_{l_{\mathbf{A}}l_{\mathbf{B}}l_{\mathbf{C}}}$, as given in Table 3, we construct

the multinomial log-likelihood by

$$\ln L(\mathbf{\Omega}|\mathbf{M}) \propto \sum_{l_{\mathbf{A}}=aa}^{AA} \sum_{l_{\mathbf{B}}=bb}^{BB} \sum_{l_{\mathbf{C}}=cc}^{CC} n_{l_{\mathbf{A}} l_{\mathbf{B}} l_{\mathbf{C}}} \ln P_{l_{\mathbf{A}} l_{\mathbf{B}} l_{\mathbf{C}}}, \tag{11}$$

from which the EM algorithm is derived to estimate haplotype frequencies, i.e.,

$$\begin{aligned}
\hat{p}_{ABC} &= \frac{1}{2n}(2n_{AABBCC} + n_{AABBCc} + n_{AABbCC} + n_{AaBBCC} \\
&\quad + \phi_1 n_{AABbCc} + \phi_2 n_{AaBBCc} + \phi_3 n_{AaBbCC} + \phi_4 n_{AaBbCc}) \\
\hat{p}_{ABc} &= \frac{1}{2n}(2n_{AABBcc} + n_{AABBCc} + n_{AABbcc} + n_{AaBBcc} \\
&\quad + \bar{\phi}_1 n_{AABbCc} + \bar{\phi}_2 n_{AaBBCc} + \phi''_4 n_{AaBbCc} + \phi_5 n_{AaBbcc}) \\
\hat{p}_{AbC} &= \frac{1}{2n}(2n_{AAbbCC} + n_{AABbCC} + n_{AAbbCc} + n_{AabbCC} \\
&\quad + \bar{\phi}_1 n_{AABbCc} + \bar{\phi}_3 n_{AaBbCC} + \phi'_4 n_{AaBbCc} + \phi_6 n_{AabbCc}) \\
\hat{p}_{Abc} &= \frac{1}{2n}(2n_{AAbbcc} + n_{AABbcc} + n_{AAbbCc} + n_{Aabbcc} \\
&\quad + \phi_1 n_{AABbCc} + \phi'''_4 n_{AaBbCc} + \bar{\phi}_5 n_{AaBbcc} + \bar{\phi}_6 n_{AabbCc}) \\
\hat{p}_{aBC} &= \frac{1}{2n}(2n_{aaBBCC} + n_{AaBBCC} + n_{aaBbCC} + n_{aaBBCc} \\
&\quad + \bar{\phi}_2 n_{AaBBCc} + \bar{\phi}_3 n_{AaBbCC} + \phi'''_4 n_{AaBbCc} + \phi_7 n_{aaBbCc}) \\
\\
\hat{p}_{aBc} &= \frac{1}{2n}(2n_{aaBBcc} + n_{AaBBcc} + n_{aaBbcc} + n_{aaBBCc} \\
&\quad + \phi_2 n_{AaBBCc} + \phi'_4 n_{AaBbCc} + \bar{\phi}_5 n_{AaBbcc} + \bar{\phi}_7 n_{aaBbCc}) \\
\hat{p}_{abC} &= \frac{1}{2n}(2n_{aabbCC} + n_{AabbCC} + n_{aaBbCC} + n_{aabbCc} \\
&\quad + \phi_3 n_{AaBbCC} + \phi''_4 n_{AaBbCc} + \bar{\phi}_6 n_{AabbCc} + \bar{\phi}_7 n_{aaBbCc}) \\
\hat{p}_{abc} &= \frac{1}{2n}(2n_{aabbcc} + n_{Aabbcc} + n_{aaBbcc} + n_{aabbCc} \\
&\quad + \phi_4 n_{AaBbCc} + \phi_5 n_{AaBbcc} + \phi_6 n_{AabbCc} + \phi_7 n_{aaBbCc})
\end{aligned} \tag{12}$$

with $\bar{\phi} = 1 - \phi$, where

$$
\begin{aligned}
\phi_1 &= \frac{p_{ABC}p_{Abc}}{p_{ABC}p_{Abc} + p_{AbC}p_{ABc}}, && \text{for genotype } AABbCc \\
\phi_2 &= \frac{p_{ABC}p_{aBc}}{p_{ABC}p_{aBc} + p_{aBC}p_{ABc}}, && \text{for genotype } AaBBCc \\
\phi_3 &= \frac{p_{ABC}p_{abC}}{p_{ABC}p_{abC} + p_{AbC}p_{aBC}}, && \text{for genotype } AaBbCC \\
\phi_4 &= \frac{p_{ABC}p_{abc}}{p_{ABC}p_{abc} + p_{AbC}p_{aBc} + p_{ABc}p_{abC} + p_{Abc}p_{aBC}}, && \text{for genotype } AaBbCc \\
\phi_4' &= \frac{p_{AbC}p_{aBc}}{p_{ABC}p_{abc} + p_{AbC}p_{aBc} + p_{ABc}p_{abC} + p_{Abc}p_{aBC}}, && \text{for genotype } AaBbCc \\
\phi_4'' &= \frac{p_{ABc}p_{abC}}{p_{ABC}p_{abc} + p_{AbC}p_{aBc} + p_{ABc}p_{abC} + p_{Abc}p_{aBC}}, && \text{for genotype } AaBbCc \\
\phi_4''' &= \frac{p_{Abc}p_{aBC}}{p_{ABC}p_{abc} + p_{AbC}p_{aBc} + p_{ABc}p_{abC} + p_{Abc}p_{aBC}}, && \text{for genotype } AaBbCc \\
\phi_5 &= \frac{p_{ABc}p_{abc}}{p_{ABc}p_{abc} + p_{Abc}p_{aBc}}, && \text{for genotype } AaBbcc \\
\phi_6 &= \frac{p_{AbC}p_{abc}}{p_{AbC}p_{abc} + p_{abC}p_{Abc}}, && \text{for genotype } AabbCc \\
\phi_7 &= \frac{p_{aBC}p_{abc}}{p_{aBC}p_{abc} + p_{abC}p_{aBc}}, && \text{for genotype } aaBbCc
\end{aligned}
\tag{13}
$$

Equations 13 and 12 construct a loop for iterations in the EM algorithm. Started with initiate values, iterations are undertaken until converged estimates are obtained. The allele frequencies and linkage disequilibrium of different orders can be estimated from the MLEs of haplotype frequencies by

$$
\begin{aligned}
\widehat{p}_1 &= \widehat{p}_{ABC} + \widehat{p}_{ABc} + \widehat{p}_{AbC} + \widehat{p}_{Abc} \\
\widehat{p}_2 &= \widehat{p}_{ABC} + \widehat{p}_{ABc} + \widehat{p}_{aBC} + \widehat{p}_{aBc} \\
\widehat{p}_3 &= \widehat{p}_{ABC} + \widehat{p}_{AbC} + \widehat{p}_{aBC} + \widehat{p}_{abC} \\
\widehat{D}_{12} &= (\widehat{p}_{ABC} + \widehat{p}_{ABc})(\widehat{p}_{abC} + \widehat{p}_{abc}) - (\widehat{p}_{AbC} + \widehat{p}_{Abc})(\widehat{p}_{aBC} + \widehat{p}_{aBc}) \\
\widehat{D}_{13} &= (\widehat{p}_{ABC} + \widehat{p}_{AbC})(\widehat{p}_{aBc} + \widehat{p}_{abc}) - (\widehat{p}_{ABc} + \widehat{p}_{Abc})(\widehat{p}_{aBC} + \widehat{p}_{abC}) \\
\widehat{D}_{23} &= (\widehat{p}_{ABC} + \widehat{p}_{aBC})(\widehat{p}_{Abc} + \widehat{p}_{abc}) - (\widehat{p}_{ABc} + \widehat{p}_{aBc})(\widehat{p}_{AbC} + \widehat{p}_{abC}) \\
\widehat{D}_{123} &= (\widehat{p}_{ABC}\widehat{p}_{Abc} + \widehat{p}_{aBc}\widehat{p}_{abC}) - (\widehat{p}_{abc}\widehat{p}_{aBC} + \widehat{p}_{ABc}\widehat{p}_{AbC})
\end{aligned}
\tag{14}
$$

The sampling variances for the MLEs of haplotype frequencies (and therefore allele frequencies and linkage disequilibria of different orders) in a three-locus analysis can be calculated from the Fisher information matrix.

The linkage disequilibria of different orders can be tested in general or individually. The

existence of linkage disequilibria is tested using the following hypotheses:

$$\begin{cases} H_0: & D_{\mathbf{AB}} = D_{\mathbf{AC}} = D_{\mathbf{BC}} = D_{\mathbf{ABC}} = 0 \\ H_1: & \text{At least one of the equalities above does not hold} \end{cases} \tag{15}$$

The LR test statistic for the significance of LD is calculated by comparing the likelihood values under the H_1 (full model) and H_0 (reduced model) using an equation similar to (7). The LR is considered to asymptotically follow a χ^2 distribution with four degrees of freedom. The MLEs of allelic frequencies under the H_0 can be estimated using the EM algorithm described above, but with the constraints

$$\begin{aligned} &(p_{ABC} + p_{ABc})(p_{abC} + p_{abc}) = (p_{AbC} + p_{Abc})(p_{aBC} + p_{aBc}), && \text{for } D_{\mathbf{AB}} \\ &(p_{ABC} + p_{AbC})(p_{aBc} + p_{abc}) = (p_{ABc} + p_{Abc})(p_{aBC} + p_{abC}), && \text{for } D_{\mathbf{AC}} \\ &(p_{ABC} + p_{aBC})(p_{Abc} + p_{abc}) = (p_{ABc} + p_{aBc})(p_{AbC} + p_{abC}), && \text{for } D_{\mathbf{BC}} \\ &p_{ABC}p_{Abc} + p_{aBc}p_{abC} = p_{aBC}p_{abc} + p_{ABc}p_{AbC}, && \text{for } D_{\mathbf{ABC}} \end{aligned} \tag{16}$$

The Dominant Model: Like two-locus linkage disequilibrium analysis, we derive the EM algorithm to estimate the haplotype frequencies for three-locus linkage disequilibrium analysis of dominant markers. When all the three markers are dominant, expected genotype frequencies for codominant markers listed in Table 3 are collapsed to form simpler Table 4. We use $n_{k_{\mathbf{A}}/k_{\mathbf{B}}/k_{\mathbf{C}}}$ to denote the observed number of observations of a two-dominant-marker genotype $k_{\mathbf{A}}/k_{\mathbf{B}}/k_{\mathbf{C}}$ $[k_{\mathbf{A}} = A\_ \ (1), oo \ (0); k_{\mathbf{B}} = B\_ \ (1), oo \ (0); k_{\mathbf{C}} = C\_ \ (1), oo \ (0)]$. In this

case, haplotype frequencies are estimated by

$$
\begin{aligned}
\widehat{p}_{ABC} &= \frac{1}{2n}(\phi_1 n_{1/1/1}) \\
\widehat{p}_{ABc} &= \frac{1}{2n}(\phi_2 n_{1/1/1} + \phi_3 n_{1/1/0}) \\
\widehat{p}_{AbC} &= \frac{1}{2n}(\phi_4 n_{1/1/1} + \phi_5 n_{1/0/1}) \\
\widehat{p}_{Abc} &= \frac{1}{2n}(\phi_6 n_{1/1/1} + \phi_7 n_{1/0/} + \phi_8 n_{ABc} + \phi_9 n_{1/0/0}) \\
\widehat{p}_{aBC} &= \frac{1}{2n}(\phi_{10} n_{1/1/1} + \phi_{11} n_{0/1/1}) \\
\widehat{p}_{aBc} &= \frac{1}{2n}(\phi_{12} n_{1/1/1} + \phi_{13} n_{0/1/1} + \phi_{14} n_{1/1/0} + \phi_{15} n_{0/1/0}) \\
\widehat{p}_{abC} &= \frac{1}{2n}(\phi_{16} n_{1/1/1} + \phi_{17} n_{1/0/1} + \phi_{18} n_{0/1/1} + \phi_{19} n_{0/0/1}) \\
\widehat{p}_{abc} &= \frac{1}{2n}(\phi_{20} n_{1/1/1} + \phi_{21} n_{1/1/0} + \phi_{22} n_{1/0/1} + \phi_{23} n_{1/0/0} + \phi_{24} n_{0/1/1} \\
&\quad + \phi_{25} n_{0/1/0} + \phi_{26} n_{0/0/1} + 2n_{0/0/0})
\end{aligned}
\tag{17}
$$

where

$$
\begin{aligned}
\phi_1 &= \frac{2p_{ABC}}{2(p_{ABC} + p_{ABc}p_{AbC} + p_{aBC}p_{AbC} + p_{ABc}p_{aBC} + p_{ABc}p_{abC} + p_{aBc}p_{AbC} + p_{aBC}p_{Abc}) - p_{ABC}^2} \\
\phi_2 &= \frac{2p_{ABc}(p_{ABC} + p_{AbC} + p_{aBC} + p_{abC})}{2(p_{ABC} + p_{ABc}p_{AbC} + p_{aBC}p_{AbC} + p_{ABc}p_{aBC} + p_{ABc}p_{abC} + p_{aBc}p_{AbC} + p_{aBC}p_{Abc}) - p_{ABC}^2} \\
\phi_3 &= \frac{2p_{ABc}(p_{ABc} + p_{Abc} + p_{aBc} + p_{abc})}{2p_{ABc}(p_{ABc} + p_{Abc} + p_{aBc} + p_{abc}) + 2p_{Abc}p_{aBc} - p_{ABc}^2} \\
\phi_4 &= \frac{2p_{AbC}(p_{ABC} + p_{ABc} + p_{aBC} + p_{aBc})}{2(p_{ABC} + p_{ABc}p_{AbC} + p_{aBC}p_{AbC} + p_{ABc}p_{aBC} + p_{ABc}p_{abC} + p_{aBc}p_{AbC} + p_{aBC}p_{Abc}) - p_{ABC}^2} \\
\phi_5 &= \frac{2p_{AbC}(p_{AbC} + p_{Abc} + p_{abC} + p_{abc})}{2p_{AbC}(p_{AbC} + p_{Abc} + p_{abC} + p_{abc}) + 2p_{Abc}p_{abC} - p_{AbC}^2} \\
\phi_6 &= \frac{2p_{Abc}(p_{ABC} + p_{aBC})}{2(p_{ABC} + p_{ABc}p_{AbC} + p_{aBC}p_{AbC} + p_{ABc}p_{aBC} + p_{ABc}p_{abC} + p_{aBc}p_{AbC} + p_{aBC}p_{Abc}) - p_{ABC}^2} \\
\phi_7 &= \frac{2p_{Abc}(p_{AbC} + p_{abC})}{2p_{AbC}(p_{AbC} + p_{Abc} + p_{abC} + p_{abc}) + 2p_{Abc}p_{abC} - p_{AbC}^2} \\
\phi_8 &= \frac{2p_{Abc}(p_{ABc} + p_{aBc})}{2p_{ABc}(p_{ABc} + p_{Abc} + p_{aBc} + p_{abc}) + 2p_{Abc}p_{aBc} - p_{ABc}^2}
\end{aligned}
\tag{18}
$$

$$
\begin{aligned}
\phi_9 &= \frac{2p_{Abc} + 2p_{abc}}{p_{Abc} + 2p_{abc}} \\
\phi_{10} &= \frac{2p_{aBC}(p_{ABC} + p_{ABc} + p_{AbC} + p_{Abc})}{2(p_{ABC} + p_{ABc}p_{AbC} + p_{aBC}p_{AbC} + p_{ABc}p_{aBC} + p_{ABc}p_{abC} + p_{aBc}p_{AbC} + p_{aBC}p_{Abc}) - p_{ABC}^2} \\
\phi_{11} &= \frac{2p_{aBC}(p_{aBC} + p_{aBc} + p_{abC} + p_{abc})}{2p_{aBC}(p_{aBC} + p_{aBc} + p_{abC} + p_{abc}) + 2p_{aBc}p_{abC} - p_{aBC}^2} \\
\phi_{12} &= \frac{2p_{aBc}(p_{ABC} + p_{AbC})}{2(p_{ABC} + p_{ABc}p_{AbC} + p_{aBC}p_{AbC} + p_{ABc}p_{aBC} + p_{ABc}p_{abC} + p_{aBc}p_{AbC} + p_{aBC}p_{Abc}) - p_{ABC}^2} \\
\phi_{13} &= \frac{2p_{aBc}(p_{aBC} + p_{abC})}{2p_{aBC}(p_{aBC} + p_{aBc} + p_{abC} + p_{abc}) + 2p_{aBc}p_{abC} - p_{aBC}^2} \\
\phi_{14} &= \frac{2p_{aBc}(p_{ABc} + p_{Abc})}{2p_{ABc}(p_{ABc} + p_{Abc} + p_{aBc} + p_{abc}) + 2p_{Abc}p_{aBc} - p_{ABc}^2} \\
\phi_{15} &= \frac{2p_{aBc} + 2p_{abc}}{p_{aBc} + 2p_{abc}} \\
\phi_{16} &= \frac{2p_{abC}(p_{ABC} + p_{ABc})}{2(p_{ABC} + p_{ABc}p_{AbC} + p_{aBC}p_{AbC} + p_{ABc}p_{aBC} + p_{ABc}p_{abC} + p_{aBc}p_{AbC} + p_{aBC}p_{Abc}) - p_{ABC}^2} \\
\phi_{17} &= \frac{2p_{abC}(p_{AbC} + p_{Abc})}{2p_{AbC}(p_{AbC} + p_{Abc} + p_{abC} + p_{abc}) + 2p_{Abc}p_{abC} - p_{AbC}^2} \\
\phi_{18} &= \frac{2p_{abC}(p_{aBC} + p_{aBc})}{2p_{aBC}(p_{aBC} + p_{aBc} + p_{abC} + p_{abc}) + 2p_{aBc}p_{abC} - p_{aBC}^2} \\
\phi_{19} &= \frac{2p_{abC} + 2p_{abc}}{p_{abC} + 2p_{abc}} \\
\phi_{20} &= \frac{2p_{ABC}p_{abc}}{2(p_{ABC} + p_{ABc}p_{AbC} + p_{aBC}p_{AbC} + p_{ABc}p_{aBC} + p_{ABc}p_{abC} + p_{aBc}p_{AbC} + p_{aBC}p_{Abc}) - p_{ABC}^2} \\
\phi_{21} &= \frac{2p_{ABc}p_{abc}}{2p_{ABc}(p_{ABc} + p_{Abc} + p_{aBc} + p_{abc}) + 2p_{Abc}p_{aBc} - p_{ABc}^2} \\
\phi_{22} &= \frac{2p_{AbC}p_{abc}}{2p_{AbC}(p_{AbC} + p_{Abc} + p_{abC} + p_{abc}) + 2p_{Abc}p_{abC} - p_{AbC}^2} \\
\phi_{23} &= \frac{2p_{abc}}{p_{Abc} + 2p_{abc}} \\
\phi_{24} &= \frac{2p_{aBC}p_{abc}}{2p_{aBC}(p_{aBC} + p_{aBc} + p_{abC} + p_{abc}) + 2p_{aBc}p_{abC} - p_{aBC}^2} \\
\phi_{25} &= \frac{2p_{abc}}{p_{aBc} + 2p_{abc}} \\
\phi_{26} &= \frac{2p_{abc}}{p_{abC} + 2p_{abc}}
\end{aligned}
$$

Equations 18 and 17 form a loop of the EM algorithm which can provide the MLEs of haplo-

type frequencies and, therefore, allele frequencies and linkage disequilibria of different orders for three dominant markers. The hypothesis tests for the degree of linkage disequilibria can be made in a way similar to the codominant model shown by Equations 15 and 16.

MONTE CAROL SIMULATION

We perform extensive simulation studies to investigate the statistical properties of the model for estimating linkage disequilibria between different dominant markers. We consider two cases, in which markers display high and low heterozygosity, respectively. Markers of high heterozygosity are simulated by similar frequencies for the alternative alleles, say 0.5 vs. 0.5, whereas those of low heterozygosity simulated by contrast frequencies for the alternative alleles, say 0.9 vs. 0.1 (Nei 1987). In both cases, the markers are assumed to be associated with a certain linkage disequilibrium (LD). These parameters are used to simulate observations for marker genotypes for different sample sizes $n = 30$, 100, 200 and 400.

Two-locus LD Analysis: For both cases of high and low heterozygosity, LD is assumed to be 0.015. This assumed LD value corresponds to a larger normalized value (Lewinton 1964) for the two markers of lower heterozygosity than of higher heterozygosity. Table 5 gives the MLEs of the allele frequencies and linkage disequilibrium and the square roots of their mean square errors for two dominant markers. The model generally provides reasonable estimates of allele frequencies even when the sample size used is as low as 30. The accuracy and precision of the parameter estimation increase dramatically when the sample size increases to 100 for the markers of high heterozygosity, but to 200 for the markers of low heterozygosity. Regardless of the degree of heterozygosity, a small sample size cannot provide an unbiased estimate of LD. For the markers of high heterozygosity, a reasonable estimate of LD requires at least 200 individuals, whereas for the markers of low heterozygosity a doubled sample size is still not sufficient (Table 5). The simulation studies allow for the estimation of empirical power for the detection of significant LD. In general, the power is quite low for the assumed LD value, which, as expected, increases with the increase of sample size.

The same set of parameters was used to simulate codominant markers with the estimation results also tabulated in Table 5. The estimation accuracy and precision of all the parameters, particularly LD, can be better estimated for codominant than dominant markers. A small sample size of 30 would be adequate for the reasonable estimate of LD for codominant markers regardless of their heterozygosity. Compared to dominant markers, codominant markers have higher power for the detection of LD, especially when the sample size is large. Different from dominant markers, allele frequencies and LD for codominant markers can be better estimated when the markers have lower heterozygosity (Table 5).

Three-locus LD Analysis: Similar simulation schemes are used for three-locus linkage disequilibrium analysis. More heterozygous markers are simulated by assuming allele frequencies $p_A = 0.5$ and $p_B = p_C = 0.6$, whereas less heterozygous markers are simulated by assuming allele frequencies $p_A = 0.9$ and $p_B = p_C = 0.8$. In both cases, linkage disequilibria are assumed to be $D_{\mathbf{AB}} = D_{\mathbf{AC}} = D_{\mathbf{BC}} = 0.015$ and $D_{\mathbf{ABC}} = 0.010$. The results from the three-locus analysis are basically consistent with those from the two-locus analysis: For dominant markers, higher heterozygosity favors the estimation of the parameters in terms of estimation precision and power, whereas for codominant markers, lower heterozygosity favors the estimation (Table 6). Beyond two-locus analysis, three-locus analysis provides the estimation of more parameters related to linkage disequilibria and, thus, can be more informative for the understanding of population structure and organization.

Results from Tables 5 and 6 allow for the comparisons between two- and three-locus linkage disequilibrium analysis. Under the same condition, the precision of parameter estimation is similar between these two analysis approaches. But three-locus analysis displays much greater power for LD detection than two-locus analysis for codominant markers with a sample size of 100 or larger. For dominant markers, the power for LD detection can be increased only for the markers of high heterozygosity. Three-locus analysis does not improve the power for the dominant markers of low heterozygosity.

An additional simulation study was conduct to estimate Type I error rates of linkage disequilibrium analyses. Marker data were repeatedly (1000 times) simulated according to the scenarios as designed above for different sample sizes, marker heterozygosities and marker types by assuming that there is no linkage disequilibrium between the markers. It appears that sample sizes do not largely affect Type I error rates for linkage disequilibrium detection. Two- and three-locus analyses may incorrectly find significant linkage disequilibria at a similar rate (Table 7). There is more chance to obtain false positive results for markers with a high heterozygosity (7–16%) than those with a low heterozygosity (1–8%). Type I errors are slightly different between dominant and codominant markers, although the pattern of differences depends on the level of marker heterozygosity.

WORKED EXAMPLE

We used the algorithm developed to estimate and test linkage disequilibria between dominant markers in hickory trees. Hickory is an oil woody species naturally distributed in the southeastern China. DNA samples were collected from 90 trees randomly drawn from three different stands in Anhui Provinces. A total of 238 RAPD markers were genotyped, aimed to study the structure and pattern of genetic variation in the population of hickory. As a demonstration, we randomly picked up a subset of markers to test and validate our algorithm proposed. Significant linkage disequilibria were detected for many marker combinations in the hickory population (Table 8). The estimates of linkage disequilibria are broadly consistent between two- and three-locus analyses, but the latter has more chances to detect the significance of linkage disequilibria for the same marker pair than the former. This result can be validated by simulation studies in the preceding section. For example, markers 19 and 144 have no significant linkage disequilibrium according to two-locus analysis, but they were detected to be significantly associated by three-locus analysis. In some cases, the significance level of the detection of linkage disequilibria can be increased when three-locus analysis is used, compared with that from two-locus analysis. For example, the estimate of the linkage

disequilibrium between markers 16 and 53 is significant at the 5% level for two-locus analysis, but it is significant at the 0.01% level for three-locus analysis. Compared with two-locus analysis, three-locus analysis has an additional advantage in the estimation of high-order linkage disequilibria (e.g., significant $D_{\mathbf{ABC}}$'s were detected among markers 102, 31 and 227, markers 46, 53 and 196, and markers 167, 8 and 31; Table 8), which may have played an important role in shaping population structure (Nielson et al. 2004).

DISCUSSION

The use of individual markers to test for the deviation of a population from Hardy-Weinberg equilibrium has become a routine approach for the inference of the structure and evolution of the population. Linkage disequilibrium (LD) analysis based on multiple markers can provide additional information about population structure by estimating the extent and distribution of nonrandom associations throughout the genome (Stephens et al. 2001; Dawson et al. 2002; Ardlie et al. 2002). For a random mating population, the LD between two markers decays with generation in a proportion depending on the recombination fraction between the markers (Lynch and Walsh 1998). Thus, by comparing the rate of LD decay over genetic distances, the evolutionary history of a population can be inferred (Tishkoff et al. 2002; Dawson et al. 2002; Gabriel et al. 2002). Also, the rate of the LD decay as a function of generation has established a fundamental principle for the high-resolution mapping of complex traits in a population (Rafalski and Morgante 2004).

While high-throughput molecular marker techniques, such as single nucleotide polymorphisms (SNPs), have been widely used to study the population structure and evolution of humans (Stephens et al. 2001; Dawson et al. 2002), cheap dominant markers still serve as an important tool for genetic research in underrepresented species (Kuang et al. 1998). More importantly, the role of dominant markers in the characterization of biochemical functions has become more pronounced in the post-genomic era with the advent of proteomic techniques (Thiellement et al. 1999; Zivy and Vienne 2000; Consoli et al. 2002). Statistical

approaches have been derived to study the structure and organization of a population with individual dominant markers (Lynch and Milligan 1994; Zhivotosky 1999; Holsinger et al. 2002). However, the estimation of LD with multilocus dominant markers has not been well explored, although the importance of multilocus analysis in population genetic studies has been increasingly recognized (Kremer et al. 2005).

In this article, we present a statistical method for estimating and testing the LD between dominant markers by implementing the EM algorithm. With this algorithm, the frequencies of haplotypes constructed by a series of nonalleles at different markers can be precisely estimated and the estimated haplotype frequencies are further used to estimate allele frequencies and linkage disequilibria. Our presentation of the method was first based on informative codominant markers, for which the estimation of haplotype frequencies can be readily formulated, and then expanded to take into account dominant markers. Although our method was presented for two- and three-locus LD analysis, it is common to estimate linkage disequilibria among four or more markers. Perhaps, a general model should be derived that can estimate and test the LD's for an arbitrary number of codominant and dominant markers. Such a multilocus LD analysis by exploiting more information at one time has advantages for increasing the precision of parameter estimation and the power to detect significant LD. Also, more information, e.g, trigenic or higher-order LD, can be detected.

Extensive simulation studies have been performed to examine the robustness and properties of LD analysis with dominant markers under different degrees of heterozygosity and different sample sizes. The method proposed in this study has proven its power to estimate and test for the LD between dominant markers, but because of their inherent adequacy of information, the use of dominant markers to study the LD is limited under some circumstances, such as a low sample size and low heterozygosities of markers. According to the results from simulation studies, the following guidelines are recommended for dominant markers to be more effectively used in practice:

(1) For dominant markers of high heterozygosity, the sample size used for a two-locus LD analysis should be 200 or more, whereas for dominant markers of low heterozygosity, the sample size should be 400 or more;

(2) Because multi-locust analysis can improve the test power, it is always recommended to perform the LD analysis with three or more dominant markers. Furthermore, multi-locus analysis provide additional information about a web of nonrandom associations among multiple markers;

(3) Codominant markers are more powerful for the estimation and test of LD than dominant markers. It is recommended to generate a mix of dominant and codominant markers for a better characterization of the genetic structure of a population (e.g., see Yan et al. 1999).

Zhivotosky (1999) and Holsinger et al. (2002) presented an approach for estimating population structure in diploids with individual dominant DNA markers. Our method has extended the use of dominant markers to characterize linkage disequilibrium among multiple loci. Although the focus in this study is on dominant markers, the methodological principle of the model proposed can be extended to consider a complex web of polygenic LD of different orders among multiallelic markers. For outcrossing populations, multiallelic markers, such as microsatellites, are crucial for the precise characterization of population structure and diversity. Our approach for LD analysis was derived within the maximum likelihood context. Bayesian approaches that have several advantages over maximum likelihood (Holsinger et al. 2002; Fu et al. 2005) may offer greater power to understand population structure, especially when a number of markers with multiple alleles are analyzed simultaneously. The computer code to perform linkage disequilibrium analyses can be requested from the corresponding author (rwu@stat.ufl.edu).

We wish to thank the two anonymous referees for their constructive comments on this

manuscript. The preparation of this manuscript is partially supported by NSF/NIH Mathematical Biology grant (No. 0540745).

LITERATURE CITED

Ardlie, K. G., L. Kruglyak and M. Seielstad, 2002 Patterns of linkage disequilibrium in the human genome. Nat. Rev. Genet. 3: 299-309.

Bennett, J. H., 1954 On the theory of random mating. Ann. Eugen. 18: 311-317.

Consoli, L., A. Lefevre, M. Zivy, D. de Vienne and C. Damerval, 2002 QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize. Plant Mol. Biol. 48: 575-81.

Dawson, E., G. R. Abecasis, S. Bumpstead, Y. Chen et al., 2002 A first-generation linkage disequilibrium map of human chromosome 22. Nature 418: 544-548.

Fu, R. W., D. K. Dey and K. E. Holsinger, 2005 Bayesian models for the analysis of genetic structure when populations are correlated. Bioinformatics 21: 1516-1529.

Gabriel, S. B., S. F. Schaffner, H. Nyuyen, J. M. Moore et al. 2002 The structure of haplotype blocks in the human genome. Science 296: 2225-2229.

Hansson, B., H. Westerdahl, D. Hasselquist, M. Akesson and S. Bensch, 2004 Does linkage disequilibrium generate heterozygosity-fitness correlations in great reed warblers? Evolution 58: 870-879.

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

Holsinger, K. E., P. O. Lewis and D. K. Dey, 2002 A Bayesian approach to inferring population structure from dominant markers. Mol. Ecol. 11: 1157-1164.

Kremer, A., H. Caron, S. Cavers, N. Colpaert, G. Gheysen, R. Gribe, M. Lemes, A. J. Lowe, R. Margis, C. Navarro and F. Salgueiro, 2005 Monitoring genetic diversity in tropical

trees with multilocus dominant markers. Heredity 95: 274-280.

Kuang, H., T. E. Richardson, S. D. Carson and B. C. Bongarten, 1998 An allele responsible for seedling death in *Pinus radiata* D. Don. Theor. Appl. Genet. 96: 640-644.

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

Liu, T., R. J. Todhunter, Q. Lu, L. Schoettinger, H. Y. Li, R. C. Littell, S. Bliss, G. Acland, G. Lust and R. L. Wu, 2006 Extent and distribution of zygotic linkage disequilibrium in canine. Genetics 174: 439-453.

Lynch, M., and B. Milligan, 1994 Analysis of population genetic structure with RAPD markers. Mol. Ecol. 3: 91-99.

Miller, A. J., and B. A. Schaal, 2006 Domestication and the distribution of genetic variation in wild and cultivated populations of the *Mesoamerican* fruit tree *Spondias purpurea* L. (*Anacardiaceae*). Mol. Ecol. 15: 1467-1480.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nielsen, D. M., M. D. Ehm, D. V. Zaykin and B. S. Weir, 2004 Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Genetics: 1029-1040.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: Models and data. Am. J. Hum. Genet. 69: 1-14.

Rafalski, A., and M. Morgante, 2004 Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. 20: 103-111.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward and E. S. Lander, 2001 Linkage disequilibrium in the human genome. Nature 411: 199-204.

Remington, D. L., J. M. Thornsberry,Y. Matsuoka, L. M. Wilson, S. R. Whitt, J. Doebley,

S. Kresovich, M. M. Goodman and E. S. Buckler, IV, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. 98: 11479-11484.

Silbiger, R. N., S. A. Christ, A. C. =Leonard et al., 1998 Preliminary studies on the population genetics of the central stoneroller (*Campostoma anomalum*) from the Great Miami River Basin, Ohio. Environ. Monit. Assess. 51: 481-495.

Stephens, J. C., J. A. Schneider, D. A. Tanguay, J. Choi et al. 2001 Haplotype variation and linkage disequilibrium in 313 human genes. Science 293: 489-493.

Thiellement, H., N. Bahrman, C. Damerval, C. Plomion, M. Rossignol, V. Santonl, D. de Vienne and M. Zivy, 1999 Proteomics for genetic and physiological studies in plants. Electrophoresis 20: 2013-2026.

Tishkoff, S. A., E. Dietzsch, W. Speed et al., 1996 Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271: 1380-1387.

Tishkoff, S. A., R. Varkonyi, N. Cahinhinan et al., 2001 Haplotype diversity and linkage disequilibrium at human G6PD: Recent origin of alleles that confer malarial resistance. Science 293: 455-462.

Tishkoff, S. A., and S. M. Williams, 2002 Genetic analysis of African populations: Human evolution and complex disease. Nat. Rev. Genet. 3: 611-621.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman and M. Kuiper, 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 23: 4407-4414.

Wang, Z. H., and R. L. Wu, 2004 A statistical model for high-resolution mapping of quantitative trait loci determining human HIV-1 dynamics. Stat. Med. 23: 3033-3051.

Weiss, K. M., and A. G. Clark, 2002 Linkage disequilibrium and the mapping of complex human traits. Trends Genet. 18: 19-24.

Williams, J. G. K., A. R. Kubelki, K. J. Livak and S. V. Tingey, 1990 DNA Polymorphism amplified by arbitrary primers are useful as genetic markers. Nucleic Acid Res. 18: 6531-6535.

Yan, G., J. Romero-Severson, M. Walton, D. D. Chadee and D. W. Severson, 1999 Population genetics of the yellow fever mosquito in Trinidad: comparisons of amplified fragment length polymorphism (AFLP) and restriction fragment length polymorphism (RFLP) markers. Mol. Ecol. 8: 951-963.

Zhivotosky, L. A., 1999 Estimating population structure in diploids with multilocus dominant DNA markers. Mol. Ecol. 8: 907-913.

Zivy, M, and D. de Vienne, 2000 Proteomics: a link between genomics, genetics and physiology. Plant Mol. Biol. 44: 575-80.

Table 1: Observations and expected frequencies of nine genotypes at two codominant markers **A** and **B**

	Marker **B**			
Marker **A**	BB	Bb	bb	Marginal
AA	n_{AABB}	n_{AABb}	n_{AAbb}	
	p_{AB}^2	$2p_{AB}p_{Ab}$	p_{Ab}^2	p_A^2
Aa	n_{AaBB}	n_{AaBb}	n_{Aabb}	
	$2p_{AB}p_{aB}$	$2p_{AB}p_{ab} + 2p_{Ab}p_{aB}$	$2p_{Ab}p_{ab}$	$2p_Ap_a$
aa	n_{aaBB}	n_{aaBb}	n_{aabb}	
	p_{aB}^2	$2p_{aB}p_{ab}$	p_{ab}^2	p_a^2
Marginal	p_B^2	$2p_Bp_b$	p_b^2	1

Table 2: Observations and expected frequencies of nine genotypes at two dominant markers **A** and **B**

	Marker **B**		
Marker **A**	$B\_\_$ (1)	*oo* (0)	Marginal
$A\_\_$ (1)	$n_{1/1}$	$n_{1/0}$	
	$p_{AB}^2 + 2p_{AB}p_{Ab}+$	p_{Ab}^2+	p_A^2+
	$2p_{AB}p_{aB} + 2p_{AB}p_{ab} + 2p_{Ab}p_{aB}$	$2p_{Ab}p_{ab}$	$2p_Ap_a$
oo (0)	$n_{0/1}$	$n_{0/0}$	
	$p_{aB}^2 + 2p_{aB}p_{ab}$	p_{ab}^2	p_a^2
Marginal	$p_B^2 + 2p_Bp_b$	p_b^2	1

Note: $A\_\_ = AA + Ao$ and $B\_\_ = BB + Bb$.

Table 3: Observations and expected frequencies of nine genotypes at three codominant markers **A**, **B**, and **C**

		Marker **C**		
Marker **A**	Marker **B**	CC	Cc	cc
AA	BB	n_{AABBCC} p^2_{ABC}	n_{AABBCc} $2p_{ABC}p_{ABc}$	n_{AABBcc} p^2_{ABc}
AA	Bb	n_{AABbCC} $2p_{ABC}p_{AbC}$	n_{AABbCc} $2(p_{ABC}p_{Abc}+p_{ABc}p_{AbC})$	n_{AABbcc} $2p_{ABc}p_{Abc}$
AA	bb	n_{AAbbCC} p^2_{AbC}	n_{AAbbCc} $2p_{AbC}p_{Abc}$	n_{AAbbcc} p^2_{Abc}
Aa	BB	n_{AaBBCC} $2p_{ABC}p_{aBC}$	n_{AaBBCc} $2p_{ABC}p_{aBc}+2p_{ABc}p_{aBC}$	n_{AaBBcc} $2p_{ABc}p_{aBc}$
Aa	Bb	n_{AaBbCC} $2(p_{ABC}p_{abC}+p_{AbC}p_{aBC})$	n_{AaBbCc} $2(p_{ABC}p_{abc}+p_{ABc}p_{abC}$ $+p_{aBc}p_{AbC}+p_{aBC}p_{Abc})$	n_{AaBbcc} $2(p_{ABc}p_{abc}+p_{Abc}p_{aBc})$
Aa	bb	n_{AabbCC} $2p_{AbC}p_{abC}$	n_{AabbCc} $2p_{AbC}p_{abc}+2p_{Abc}p_{abC}$	n_{Aabbcc} $2p_{Abc}p_{abc}$
aa	BB	n_{aaBBCC} p^2_{aBC}	n_{aaBBCc} $2p_{aBC}p_{aBc}$	n_{aaBBcc} p^2_{aBc}
aa	Bb	n_{aaBbCC} $2p_{aBC}p_{abC}$	n_{aaBbCc} $2(p_{aBC}p_{abc}+p_{aBc}p_{abC})$	n_{aaBbcc} $2p_{aBc}p_{abc}$
aa	bb	n_{aabbCC} p^2_{abC}	n_{aabbCc} $2p_{abC}p_{abc}$	n_{aabbcc} p^2_{abc}

Table 4: Observations and expected frequencies of eight genotypes at three dominant markers **A**, **B**, and **C**

		Marker **C**	
Marker **A**	Marker **B**	$C\_\_$ (1)	oo (0)
$A\_\_$ (1)	$B\_\_$ (1)	$n_{1/1/1}$	$n_{1/1/0}$
		$p^2_{ABC} + 2p_{ABC}p_{ABc}$	p^2_{ABc}
		$+2p_{ABC}p_{AbC} + 2(p_{ABC}p_{Abc} + p_{ABc}p_{AbC})$	$+2p_{ABc}p_{Abc}$
		$+2p_{ABC}p_{aBC} + 2p_{ABC}p_{aBc} + 2p_{ABc}p_{aBC}$	$+2p_{ABc}p_{aBc}$
		$+2(p_{ABC}p_{abC} + p_{AbC}p_{aBC}) + 2(p_{ABC}p_{abc} + p_{ABc}p_{abC}$	$+2(p_{ABc}p_{abc} + p_{Abc}p_{aBc})$
		$+p_{aBc}p_{AbC} + p_{aBC}p_{Abc})$	
$A\_\_$ (1)	oo (0)	$n_{1/0/1}$	$n_{1/0/0}$
		$p^2_{AbC} + 2p_{AbC}p_{Abc}$	p^2_{Abc}
		$+2p_{AbC}p_{abC} + 2p_{AbC}p_{abc} + 2p_{Abc}p_{abC}$	$+2p_{Abc}p_{abc}$
oo (0)	$B\_\_$ (1)	$n_{0/1/1}$	$n_{0/1/0}$
		$p^2_{aBC} + 2p_{aBC}p_{aBc}$	p^2_{aBc}
		$+2p_{aBC}p_{abC} + 2(p_{aBC}p_{abc} + p_{aBc}p_{abC})$	$+2p_{aBc}p_{abc}$
oo (0)	oo (0)	$n_{0/0/1}$	$n_{0/0/0}$
		$p^2_{abC} + 2p_{abC}p_{abc}$	p^2_{abc}

Note: $A\_\_ = AA + Ao$, $B\_\_ = BB + Bo$ and $C\_\_ = CC + Co$.

Table 5: Averaged MLEs of allele frequencies and linkage disequilibrium between two markers under different levels of heterozygosity based on 1000 replicated simulations. The numbers in the parentheses are the square roots of the mean square errors of the MLEs.

	High heterozygosity				Low heterozygosity			
n	$p_A = 0.5$	$p_B = 0.6$	$D = 0.015$	Power	$p_A = 0.9$	$p_B = 0.8$	$D = 0.015$	Power
				Dominant markers				
30	0.5139	0.6118	-0.0071	12	0.9524	0.8371	-0.0003	4
	(0.0821)	(0.0898)	(0.0909)		(0.0846)	(0.1168)	(0.0310)	
100	0.5037	0.6064	0.0097	12	0.9242	0.8061	-0.0021	9
	(0.0454)	(0.0472)	(0.0419)		(0.0612)	(0.0537)	(0.0310)	
200	0.5026	0.6013	0.0111	15	0.9100	0.8038	0.0010	11
	(0.0284)	(0.034)	(0.0301)		(0.0465)	(0.0348)	(0.0302)	
400	0.4997	0.5998	0.0139	19	0.9037	0.8018	0.0017	12
	(0.0196)	(0.0234)	(0.0197)		(0.0274)	(0.0244)	(0.0260)	
				Codominant markers				
30	0.4965	0.5975	0.0147	14	0.9000	0.8023	0.0139	13
	(0.0636)	(0.0617)	(0.0458)		(0.0385)	(0.0508)	(0.0236)	
100	0.4984	0.6016	0.0149	20	0.8989	0.8008	0.0143	22
	(0.037)	(0.0358)	(0.0235)		(0.0219)	(0.0282)	(0.0127)	
200	0.4978	0.5996	0.0146	24	0.9000	0.7989	0.0152	45
	(0.0240)	(0.0237)	(0.0172)		(0.0145)	(0.0220)	(0.0086)	
400	0.5005	0.6003	0.0151	37	0.8997	0.7999	0.0152	72
	(0.0182)	(0.0175)	(0.0120)		(0.0111)	(0.0141)	(0.0066)	

Note: Power is calculated as the percentage of the simulations, in which significant LD is detected at the 5% significance level, over all the simulations.

Table 6: Averaged MLEs of allele frequencies and linkage disequilibrium between three markers under different levels of heterozygosity based on 1000 replicated simulations. The numbers in the parentheses are the square roots of the mean square errors of the MLEs.

	Estimate							Power				
n	$p_A = 0.5/0.9$	$p_B = 0.6/0.8$	$p_C = 0.6/0.8$	$D_{\mathbf{AB}} = 0.015$	$D_{\mathbf{AC}} = 0.015$	$D_{\mathbf{BC}} = 0.015$	$D_{\mathbf{ABC}} = 0.01$	Overall	$D_{\mathbf{AB}}$	$D_{\mathbf{AC}}$	$D_{\mathbf{BC}}$	$D_{\mathbf{ABC}}$
Dominant markers of high heterozygosity												
30	0.4958	0.6027	0.6108	-0.0089	-0.0047	-0.0137	0.0108	17	13	12	14	10
	(0.0828)	(0.0878)	(0.0872)	(0.0897)	(0.0828)	(0.0871)	(0.0491)					
100	0.4998	0.6079	0.6079	0.0086	0.0088	0.0081	0.0227	23	31	36	32	22
	(0.0445)	(0.0502)	(0.0463)	(0.0432)	(0.0438)	(0.0470)	(0.0425)					
200	0.5000	0.5977	0.6002	0.0113	0.0152	0.0133	0.0202	32	37	38	37	23
	(0.0297)	(0.0322)	(0.0304)	(0.0269)	(0.0289)	(0.0305)	(0.0328)					
400	0.4972	0.6039	0.5973	0.0146	0.0147	0.0144	0.0107	39	47	47	51	27
	(0.0213)	(0.024)	(0.0212)	(0.0202)	(0.0200)	(0.0198)	(0.0201)					
Dominant markers of low heterozygosity												
30	0.9467	0.8332	0.8341	-0.0011	-0.0040	-0.0088	-0.0078	7	3	3	8	7
	(0.0873)	(0.1165)	(0.1155)	(0.0321)	(0.0278)	(0.0498)	(0.0468)					
100	0.9188	0.8055	0.8048	-0.0018	-0.0038	-0.0086	-0.0069	8	11	11	17	13
	(0.0605)	(0.0554)	(0.0504)	(0.0323)	(0.0300)	(0.0444)	(0.0402)					
200	0.9087	0.8021	0.8035	0.0006	-0.0029	-0.0015	0.0009	10	23	23	21	18
	(0.0423)	(0.0353)	(0.0374)	(0.0296)	(0.0279)	(0.0395)	(0.0359)					
400	0.9035	0.8013	0.8022	0.0019	0.0031	0.0054	0.0089	16	34	33	28	22
	(0.0265)	(0.0246)	(0.0259)	(0.0250)	(0.0255)	(0.0306)	(0.0287)					
Codominant markers of high heterozygosity												
30	0.5038	0.5968	0.5930	0.0138	0.0103	0.0112	0.0097	19	33	36	37	31
	(0.0637)	(0.0614)	(0.0631)	(0.0475)	(0.0473)	(0.0450)	(0.0310)					
100	0.5001	0.6010	0.5974	0.0149	0.0154	0.0127	0.0106	34	38	39	47	33
	(0.0342)	(0.0363)	(0.0346)	(0.0251)	(0.0250)	(0.0231)	(0.0167)					
200	0.4988	0.5990	0.6014	0.0163	0.0158	0.0140	0.0099	57	45	44	57	41
	(0.0242)	(0.0245)	(0.0237)	(0.0174)	(0.0182)	(0.0173)	(0.0118)					
400	0.5003	0.5983	0.6010	0.0147	0.0150	0.0144	0.0093	75	55	59	67	53
	(0.0186)	(0.0175)	(0.0176)	(0.0126)	(0.0127)	(0.0120)	(0.0084)					
Codominant markers of low heterozygosity												
30	0.8934	0.7941	0.7963	0.0166	0.0202	0.0156	0.0191	16	31	34	20	22
	(0.0379)	(0.0504)	(0.0525)	(0.0222)	(0.0229)	(0.0310)	(0.0271)					
100	0.9010	0.8006	0.7973	0.0155	0.0163	0.0148	0.0198	58	75	71	52	64
	(0.0197)	(0.0293)	(0.0295)	(0.0121)	(0.0122)	(0.0162)	(0.0143)					
200	0.8993	0.7993	0.8016	0.0159	0.0161	0.0148	0.0207	92	88	91	64	86
	(0.0146)	(0.0202)	(0.0191)	(0.0086)	(0.0087)	(0.0107)	(0.0094)					
400	0.8997	0.7994	0.8004	0.0151	0.0155	0.0149	0.0214	100	96	98	84	98
	(0.0106)	(0.0142)	(0.0150)	(0.0063)	(0.0063)	(0.0085)	(0.0076)					

Note: Power is calculated as the percentage of the simulations, in which significant LD is detected at the 5% significance level, over all the simulations.

Table 7: Type I error rates of two- and three-locus linkage disequilibrium analysis

	High Heterozygosity		Low Heterozygosity	
n	Dominant	Codominant	Dominant	Codominant
Two-locus analysis				
30	13	11	2	6
100	12	7	3	4
200	9	8	3	6
400	10	7	3	5
Three-locus analysis				
30	16	8	2	8
100	14	8	2	6
200	14	10	1	3
400	16	13	2	5

Note: Type I errors for the three-locus analysis were estimated from the simulated data containing no linkage disequilibria of any order.

Table 8: Estimates of linkage disequilibria and their log-likelihood ratio (LR) test statistics for a small subset of RAPD markers genotyped to study the population structure of hickory trees by three- and two-locus analyses

No. Marker			Three-locus Analysis								Two-locus Analysis					
A	**B**	**C**	$D_{\mathbf{AB}}$	$D_{\mathbf{AC}}$	$D_{\mathbf{BC}}$	$D_{\mathbf{ABC}}$	$\mathrm{LR}_{\mathbf{AB}}$	$\mathrm{LR}_{\mathbf{AC}}$	$\mathrm{LR}_{\mathbf{BC}}$	$\mathrm{LR}_{\mathbf{ABC}}$	$D_{\mathbf{AB}}$	$D_{\mathbf{AC}}$	$D_{\mathbf{BC}}$	$\mathrm{LR}_{\mathbf{AB}}$	$\mathrm{LR}_{\mathbf{AC}}$	$\mathrm{LR}_{\mathbf{BC}}$
140	19	144	0.0000	0.0000	0.0334	0.0333	0.00	0.00	20.99	0.85	0.0000	0.0000	0.0334	0.00	0.00	0.84
208	184	166	-0.0445	-0.0122	-0.0556	0.0307	1.76	0.46	2.89	0.25	-0.0445	-0.0122	-0.0556	1.18	0.15	1.52
238	43	163	-0.0512	-0.0248	-0.0022	0.0713	4.45	2.12	0.01	1.91	-0.0513	-0.0247	-0.0020	2.33	0.49	0.00
102	31	227	-0.0029	0.1066	0.0530	-0.0238	0.04	16.03	8.71	14.18	-0.0029	0.1066	0.0530	0.00	10.17	2.23
113	225	144	-0.0521	0.0225	0.0163	-0.0247	6.85	1.80	0.55	1.49	-0.0521	0.0225	0.0163	2.62	0.54	0.24
139	124	227	-0.1524	-0.0987	0.0121	-0.0371	9.92	4.76	0.19	0.78	-0.1531	-0.0993	0.0121	5.42	2.11	0.10
167	8	31	-0.0158	0.0219	0.0279	-0.0377	0.83	0.49	3.99	4.11	-0.0158	0.0219	0.0279	0.14	0.35	0.46
46	53	196	0.0704	-0.0222	0.0401	-0.0530	17.69	0.64	6.56	9.55	0.0704	-0.0223	0.0401	4.95	0.36	1.49