Statistical Applications in Genetics and Molecular Biology

Volume 9, *Issue* 1 2010 *Article* 16

Trilocus Disequilibrium Analysis of Multiallelic Markers in Outcrossing Populations

Arthur Berg, *Pennsylvania State University and Beijing Forestry University*
Qiuling He, *Nanjing Forestry University*
Ye Shen, *Nanjing Forestry University*
Ying Chen, *Nanjing Forestry University*
Minren Huang, *Nanjing Forestry University*
Rongling Wu, *Pennsylvania State University and Beijing Forestry University*

Recommended Citation:
Berg, Arthur; He, Qiuling; Shen, Ye; Chen, Ying; Huang, Minren; and Wu, Rongling (2010) "Trilocus Disequilibrium Analysis of Multiallelic Markers in Outcrossing Populations," *Statistical Applications in Genetics and Molecular Biology*: Vol. 9: Iss. 1, Article 16.
DOI: 10.2202/1544-6115.1528

Trilocus Disequilibrium Analysis of Multiallelic Markers in Outcrossing Populations

Arthur Berg, Qiuling He, Ye Shen, Ying Chen, Minren Huang, and Rongling Wu

Abstract

Multiallelic markers, such as microsatellites, provide a powerful tool for studying the genetic structure and organization of an outcrossing population. However, statistical methods of analyzing multiallelic markers in current literature are limited in scope due to the complexity of the multiple alleles. We present a closed-form EM algorithm framework to estimate trigenic linkage disequilibria coefficients of three multiallelic markers and present joint and separate statistical hypothesis tests of different linkage disequilibria. Linkage disequilibria analysis with three multiallelic markers is shown to be considerably more powerful than a two marker analysis or a three marker analysis that treats the multiallelic markers as biallelic markers. A three multiallelic marker model was used to analyze marker data from *Lycoris longituba*, a tulip-like ornamental plant in China, where each marker consisted of two to four distinct alleles. This algorithm will be useful for studying the pattern of genetic variation for outcrossing populations.

KEYWORDS: EM algorithm, linkage disequilibrium, multiallelic marker, natural population, trigenic disequilibrium

Author Notes: The preparation of this manuscript is partially supported by NSF/NIH Mathematical Biology grant (No. 0540745) and the Changjiang Scholars Award at Beijing Forestry University. We also wish give thanks to an anonymous referee who provided several useful suggestions that have considerably improved this manuscript.

INTRODUCTION

Molecular markers have been served as a useful tool to study the genetic structure and diversity of a population. By testing Hardy-Weinberg equilibrium and gametic linkage disequilibrium for different markers, the genetic properties of a population can be quantified. Widely used marker systems in current studies include biallelic markers, such as restricted fragment length polymorphisms and single nucleotide polymorphisms, dominant markers, such as random amplified polymorphic DNA, amplified fragment length polymorphisms, and proteomic markers, and multiallelic markers, such as microsatellites. Most early statistical methods for disequilibrium analysis are based on biallelic codominant markers (Weir and Ott 1996). Slatkin and Excoffier (1996) published a seminal paper for implementing the EM algorithm to estimate haplotype frequencies for this type of markers. More recently, the analysis and modeling of dominant DNA markers has received considerable attention in the past decade (Zhivotovsky 1999; Miller and Schaal 2006). Li *et al.* (2007) developed a general framework for disequilibrium analysis of biallelic and dominant markers in a diploid outcrossing population.

Linkage disequilibrium (LD) analysis is an important method in population genetics as described in the classical papers Bennett (1954), Lewontin and Kojima (1960), and Hill (1974), and modern usages of linkage disequilibrium has been nicely surveyed in a recent *Nature Reviews Genetics* article by Slatkin (2008). LD for multiple loci was developed with a probabilistic model by Geiringer (1944) in the *Annals of Mathematical Statistics*, and also provided an algorithmic framework to estimate the higher-order linkage coefficients. By adopting an alternative definition of higher-order disequilibrium given in Dausset *et al.* (1978), Gorelick and Laubichler (2004) presented an alternative model for multilocus linkage disequilibrium. These models and algorithms, however, have been limited to biallelic and dominant markers, and here we present a general theory for higher-order LD estimation with multiallelic markers.

In many situations, biallelic and dominant markers cannot adequately reflect the highly polymorphic nature of an outcrossing population which is still in a wild or semi-wild status. Microsatellites or simple sequence repeats that consist of repeating units of 1-6 base pairs in length have power to detect multiple alleles at a single locus and have been increasingly used to understand the genetic diversity of outcrossing populations. However, because of increasing complexity, statistical approaches for analyzing multiallelic microsatellite markers have not well ben developed. Some studies simply collapse three or more alleles into two groups, one containing the most prevalent alleles and

the other containing all other alleles, and then a biallelic analysis is performed (Long *et al.* 1995). This collapsing step can dramatically change the power of the linkage disequilibrium analysis as indicated in (Weir and Cockerham 1978). Other studies attempt to take into account the multiallelic characteristic of microsatellites (Mohlke *et al.* 2001), but genetic theories for marker analysis in these studies are limited to one or two multiallelic markers. Analyzing individual multiallelic markers for Hardy-Weinberg equilibrium has been extensively studied (Hernandez and Weir 1989; Guo and Thompson 1992; Louis and Dempster 1987). Analyzing pairs of multiallelic markers for linkage disequilibrium using the EM algorithm has been proposed in (Kalinowski and Hedrick 2001). Zaykin *et al.* (2008) show a correlation-based approach for inferencing and testing linkage disequilibrium with multiple alleles. Kim *et al.* (2008) consider a measure of multilocus LD with an arbitrary number of loci in a biallelic setting by utilizing a multiple order Markov chain model. Coalescence theory provides an alternative approach to defining LD, requiring the modeling of ancestry, yet can efficiently handle an arbitrary number of markers (cf. Hössjer *et al.* (2009)).

It has been recognized that three-marker analysis is advantageous in the detection power of disequilibria and their estimation precision than two-marker analysis because of more data used in the former (Li *et al.* 2007; Li and Wu 2009). However, this advantage has never been justified for the analysis of multiallelic markers in which an exponentially increasing number of disequilibria will need to be estimated although more data are involved. The motivation of this article is to present a detailed analysis of linkage disequilibria with three multiallelic markers, each with any number of alleles in outcrossing populations, and also provide an analytical procedure for estimating and testing trigenic linkage disequilibria at different orders. Computer simulations were conducted and indicate that a three-locus disequilibrium analysis is not only more powerful for detecting the existence of disequilibria of different kinds, but also more precise for estimating the value of disequilibria, as compared to traditional two-locus analysis. Disequilibrium analysis with real data from a plant population genetic study validates the utilization and usefulness of the new approach.

DISEQUILIBRIUM MODEL FOR THREE MULTIALLELIC MARKERS

Consider an outcrossing population from which a sample of n unrelated individuals are drawn randomly. Microsatellites are used to genotype these

DOI: 10.2202/1544-6115.1528

sampled individuals, leading to the observation of multiallelic markers. The number of alleles at a single marker may vary from locus to locus. We consider the trigenic linkage disequilibrium estimation and testing of three multiallelic markers **A**, **B**, and **C** each in Hardy-Weinberg equilibrium.

Suppose there are m_1 alleles at marker **A**, m_2 alleles at marker **B**, and m_3 alleles at marker **C**. The general haplotype model for the $m_1m_2m_3$ haplotypes is written as

$$p_{ijk} = p_i q_j r_k + (-1)^{i+j} D_{ij} + (-1)^{i+k} D_{ik} + (-1)^{j+k} D_{jk} + (-1)^{i+j+k} D_{ijk} \quad (1)$$

(for $i = 1, \ldots, m_1; j = 1, \ldots, m_2; k = 1, \ldots, m_3$) where p_i, q_j, and r_k are allele frequencies for the i^{th}, j^{th}, and k^{th} alleles, respectively; D_{ij}, D_{jk}, and D_{ik} are the digenic linkage disequilibrium coefficients, and D_{ijk} is the trigenic linkage disequilibrium coefficient. A normalized measure of the trigenic linkage disequilibrium coefficient is discussed in Robinson *et al.* (1991). It would seem there are more linkage coefficients than equations in (3), but the disequilibrium coefficients satisfy a number of restrictions allowing the disequilibrium model in (3) to be identifiable (Weir 1979).

Let $i_1 i_2 j_1 j_2 k_1 k_2$ $(1 \le i_1 \le i_2 \le m_1, 1 \le j_1 \le j_2 \le m_2, k_1 \le k_2 \le m_3)$ denote a general genotype formed by the three multiallelic markers, and let $P_{i_1 i_2 j_1 j_2 k_1 k_2}$ and $n_{i_1 i_2 j_1 j_2 k_1 k_2}$ denote the genotype probability and observation count, respectively, for this genotype. The genotype probability can be expressed in terms of the haplotype probabilities in the following way:

$$P_{i_1 i_2 j_1 j_2 k_1 k_2} = \begin{cases} p_{ijk}^2, & i_1 = i_2 := i; \quad j_1 = j_2 := j; \quad k_1 = k_2 := k \\ 2p_{ijk_1}p_{ijk_2}, & i_1 = i_2 := i; \quad j_1 = j_2 := j; \quad k_1 \neq k_2 \\ 2p_{ij_1k}p_{ij_2k}, & i_1 = i_2 := i; \quad j_1 \neq j_2; \quad k_1 = k_2 := k \\ 2p_{i_1jk}p_{i_2jk}, & i_1 \neq i_2; \quad j_1 = j_2 := j; \quad k_1 = k_2 := k \\ 2p_{ij_1k_1}p_{ij_2k_2} + 2p_{ij_1k_2}p_{ij_2k_1}, & i_1 = i_2 := i; \quad j_1 \neq j_2; \quad k_1 \neq k_2 \\ 2p_{i_1jk_1}p_{i_2jk_2} + 2p_{i_1jk_2}p_{i_2jk_1}, & i_1 \neq i_2; \quad j_1 = j_2 := j; \quad k_1 \neq k_2 \\ 2p_{i_1j_1k}p_{i_2j_2k} + 2p_{i_1j_2k}p_{i_2j_1k}, & i_1 \neq i_2; \quad j_1 \neq j_2; \quad k_1 = k_2 := k \\ 2p_{i_1j_1k_1}p_{i_2j_2k_2} + 2p_{i_1j_1k_2}p_{i_2j_2k_1} & \\ \quad +2p_{i_1j_2k_1}p_{i_2j_1k_2} + 2p_{i_1j_2k_2}p_{i_2j_1k_1}, & i_1 \neq i_2; \quad j_1 \neq j_2; \quad k_1 \neq k_2 \end{cases}$$

A multinomial mixture likelihood based on the above probabilities and observed genotype counts $n_{i_1 i_2 j_1 j_2 k_1 k_2}$ gives the motivation behind the haplotype frequency estimation.

Estimation of the trigenic linkage disequilibrium coefficients is detailed in Appendix C, but we provide summary of the approach here. In the case of three multiallelic markers, we first provide a procedure for haplotype frequency estimation that differs slightly from the general methodology that is discussed in Excoffier and Slatkin (1995). As in Excoffier and Slatkin (1995), we utilize the EM algorithm, but differences are present in the expectation step involving the double and triple heterozygote counts. Upon producing haplotype

frequency estimates, the disequilibrium coefficients are estimated by reparameterizing the coefficients and finally reversing the original reparameterization.

In the following section, we describe the hypothesis tests that are available by considering various hypotheses of linkage disequilibrium.

HYPOTHESIS TESTING

Overall linkage disequilibria involving three multiallelic markers can be tested by formulating the following null and alternative hypotheses:.

$$H_0 : D_{ijk} = 0, \quad \text{for all } i, j, k \quad \text{vs.} \quad H_1 : \text{ At least one equality does not hold.}$$

The general form of the log-likelihood is given by

$$\log L = \sum \hat{P}_{i_1 i_2 j_1 j_2 k_1 k_2} n_{i_1 i_2 j_1 j_2 k_1 k_2} \tag{2}$$

where the sum is over all possible genotypes $i_1 i_2 j_1 j_2 k_1 k_2$. This log-likelihood quantity can be computed under both the null and alternative hypotheses. Under the null hypothesis, $\hat{P}_{i_1 i_2 j_1 j_2 k_1 k_2}$ is calculated under the null with haplotype frequencies derived from the allele frequency estimates; i.e., $\hat{p}_{ijk} = \hat{p}_i \hat{q}_j \hat{r}_k$. The resulting log-likelihood under the null is expressed as $\log L_0$. The log-likelihood under the alternative hypothesis, $\log L_1$, is computed with (2) where $\hat{P}_{i_1 i_2 j_1 j_2 k_1 k_2}$ is calculated under the alternative hypothesis with haplotype frequencies derived from the EM algorithm described in Appendix C. The likelihood ratio is then calculated by

$$\text{LR} = -2(\log L_0 - \log L_1)$$

which is asymptotically χ^2-distributed with $m_1 m_2 m_3 - m_1 - m_2 - m_3 + 2$ degrees of freedom.

We may also be interested in testing the significance of individual trigenic linkage disequilibria with a null hypothesis generally expressed as

$$H_0 : D_{ijk} = 0 \quad \text{vs.} \quad H_1 : \ D_{ijk} \neq 0, \ \text{for some } i, j, k.$$

Once the digenic and trigenic linkage disequilibrium coefficients are computed, the log-likelihood under the null again follows (2) where $\hat{P}_{i_1 i_2 j_1 j_2 k_1 k_2}$ is calculated under the following constraints

$$\hat{p}_{ijk} = \hat{p}_i \hat{q}_j \hat{r}_k + (-1)^{i+j} \hat{D}_{ij} + (-1)^{i+k} \hat{D}_{ik} + (-1)^{j+k} \hat{D}_{jk}$$

$$\hat{p}_{i'j'k'} = 1 - \sum_{(i'',j'',k'') \neq (i'j'k')} \hat{p}_{i''j''k''} \quad \text{for any single choice of } (i', j', k') \neq (i, j, k)$$

DOI: 10.2202/1544-6115.1528

The second constraint forces the sum of the haplotype frequency estimates to equal one. The log-likelihood under the alternative is computed the same as above. The likelihood ratio statistic under this hypothesis is asymptotically χ^2-distributed with one degree of freedom.

Finally, we may be interested in testing where a given allele A_{i_0} at marker **A** has significant overall disequilibria with all possible alleles at markers **B** and **C**. This suggests the hypothesis test

$$H_0 : D_{i_0 j} = 0, D_{i_0 k} = 0, \text{and } D_{i_0 jk} = 0 \text{ for every } j, k$$
$$\text{vs.}$$
$$H_1 : \text{ At least one equality does not hold}$$

The haplotype frequencies under the null hypothesis can be estimated under the constraints

$$\hat{p}_{i_0 jk} = \hat{p}_{i_0}\hat{q}_j\hat{r}_k \qquad \text{for every } j, k$$
$$\hat{p}_{i_1 j' k'} = 1 - \sum_{(i'',j'',k'')\neq(i_1 j' k')} \hat{p}_{i'' j'' k''} \qquad \text{for any single choice of } i_1 \neq i_0, j', \text{ and } k'.$$

The likelihood ratio statistic under this hypothesis is asymptotically χ^2-distributed with $m_2 m_3 - 1$ degrees of freedom.

WORKED EXAMPLE

A population genetic project was initiated at Nanjing Forestry University, China, to explore the origin of *Lycoris longituba* and its evolutionary process. *L. longituba* is a perennial, bulbiferous, herbaceous plant of the family *Amaryllidaceae*. It is native to China, mainly distributed in the Langya Mountain of Anhui province, and the Baohua Mountain and Xuyi city of Jiangsu province. The three distribution areas are distant about 50 km from one another. *L. longituba* has a rare biological characteristic to angiosperms, i.e., its vegetative growth and reproduction are discrete. This species has also large, colorful, and fragrant flowers, making it of high ornamental value. In addition to its tremendous diversity in morphology and form, *L. longituba* can serve as ideal material for plant genetic studies.

A sample of 32 plants were randomly selected from a natural population of *L. longituba*, aimed to study the pattern of genetic diversity and evolutionary process of this species. Microsatellites show extensive length polymorphisms, in which one or a few nucleotide sequences repeat tandemly for varying times

and also the mutation frequency increases with the length of repeating sequence. In this study, a panel of 16 microsatellites were typed from this sample, producing markers with varying numbers of alleles at a locus (He *et al.* 2009). The complete dataset is provided in Appendix B.

The new model was used to analyze the linkage disequilibria of all the microsatellites genotyped. The three-locus model produces 480 combinations for disequilibrium analysis, whereas the two-locus model produces only 40 pairs. For the three-locus analysis, markers were analyzed by considering all possible alleles at each marker (the new model), and also by collapsing them into "biallelic" markers. This collapsing was performed by selecting the most frequent allele as the "first" allele and combining all other alleles into the "second" allele. Among all possible 480 three-locus combinations, only one consisted of markers 1, 3, and 17 has a significant disequilibrium based on the new model. Table 1 summarizes the results for the disequilibrium analysis of three microsatellite markers, 1, 3, and 17. These three markers have different numbers of alleles, 4, 3, and 4, respectively. There are $3\times2+3\times2+3\times3=21$ digenic linkage disequilibria and $3\times2\times3=18$ trigenic linkage disequilibria for the three-multiallelic-locus model, whereas these numbers are three and one for the two-"biallelic"-locus model. For a two-multiallelic-locus model, there are $3\times2=6$ disequilibria for markers 1 and 3, $2\times3=6$ for markers 3 and 17, and 3×3 for markers 1 and 17. The two-"biallelic" marker model did not detect any significant linkage disequilibria between these three markers, but highly significant or significant disequilibria were identified between each pair of these markers by the model that considers all alleles at each marker (Table 1). Three-locus model increases the power of disequilibrium detection. An overall linkage disequilibrium was detected to be significant ($p=0.0453$) by the three-"biallelic"-locus model and highly significant ($p=0.00018$) by the three-multiallelic-locus model. It is cautioned, however, that the 3-marker p-value may not remain significant when accounting for the 480 3-way combinations tested; an overly-conservative Bonferroni correction would yield an adjusted p-value of .0864.

DOI: 10.2202/1544-6115.1528

Table 1: Tabulated p-values for Hardy-Weinberg Disequilibria and Linkage Disequilibria with two- and three-marker analyses. The two-marker p-values are provided and listed such that the top left value corresponds to a comparison of M_1 and M_3, the bottom left value compares M_3 with M_{17}, and the right value compares M_1 with M_{17}.

ID	# of alleles	allele freq	HWD	"biallelic" 2-marker	2-marker	"biallelic" 3-marker	3-marker
M_1	4	(.33,.41,.20,.06)	.4817	.5853	3×10^{-6}		
M_3	3	(.47,.40,.13)	.0024	.4644	.0200	.04528	.00018
M_{17}	4	(.37,.27,.15,.22)	.0256	.2247	.00798		

COMPUTER SIMULATIONS

The performance and properties of the proposed method for linkage disequilibrium estimation and testing with three multiallelic markers is evaluated through simulation studies. Three markers with two, three, and four alleles and corresponding allele probabilities $p = (.75, .25)$, $q = (.3, .3, .4)$, and $r = (.5, .2, .1, .2)$ are used to simulate marker data with three sample sizes, $n = 30$, 100, and 200. All simulations were repeated over 10,000 realizations of randomly generated marker data.

In the first simulation, marker data under the null hypothesis of no linkage disequilibrium (of any order) is generated. The average parameter estimates of the allele frequencies and distinct disequilibrium coefficients with corresponding standard errors are provided in Appendix A. All parameters estimates appear unbiased with standard errors decreasing with the sample size.

Linkage disequialibria was tested with five different χ^2 statistics. The first statistic, T_1, utilizes the complete data in a three multiallelic marker test. The second statistic, T_2, collapses the three multiallelic markers into biallelic markers producing a biallelic test with three markers. This collapsing was performed by selecting the most frequent allele as the "first" allele and combining all other alleles into the "second" allele. The last three statistics, T_3, T_4 and T_5, considers each pair of the three markers and performs a multiallelic test with two markers.

Table 2: Measuring the type I error of test statistics T_1 through T_5 where marker data was simulated under no LD. Tests with measurements close to the threshold have correct level.

$\boldsymbol{n = 30}$	T_1	T_2	T_3	T_4	T_5
$\Pr[T < .05]$	.0664	.0859	.0783	.0651	.0944
$\Pr[T < .01]$	.0100	.0161	.0163	.0128	.0198
$\Pr[T < .001]$	.0008	.0015	.0016	.0011	.0018

$\boldsymbol{n = 100}$	T_1	T_2	T_3	T_4	T_5
$\Pr[T < .05]$	.1062	.0705	.0592	.0644	.0725
$\Pr[T < .01]$	.0215	.0154	.0122	.0122	.0172
$\Pr[T < .001]$	.0021	.0014	.0017	.0013	.0035

$\boldsymbol{n = 200}$	T_1	T_2	T_3	T_4	T_5
$\Pr[T < .05]$	.0883	.0565	.0524	.0565	.0617
$\Pr[T < .01]$	.0199	.0104	.0111	.0124	.0115
$\Pr[T < .001]$	.0029	.0009	.0012	.0012	.0012

Interestingly, the three multiallelic marker test statistics was closest to the nominal level at the small sample size of $n = 30$, but this is arguably an artifact of the particular choice of the allele frequency parameters. More inflated values of the type I error are observed in the simulations with larger n.

In the second simulation, marker data with linkage disequilibria is generated. The following are the linkage disequilibria coefficients used in the simulation:

$\mathcal{D}_{111}$	$\mathcal{D}_{112}$	$\mathcal{D}_{113}$	$\mathcal{D}_{114}$	$\mathcal{D}_{121}$	$\mathcal{D}_{122}$	$\mathcal{D}_{123}$	$\mathcal{D}_{124}$	$\mathcal{D}_{131}$
.01	-.02	-.01	-.04	-.02	.03	0	0	.05

$\mathcal{D}_{132}$	$\mathcal{D}_{133}$	$\mathcal{D}_{211}$	$\mathcal{D}_{212}$	$\mathcal{D}_{213}$	$\mathcal{D}_{221}$	$\mathcal{D}_{222}$	$\mathcal{D}_{223}$
-.06	.01	0	.04	0	0	-.01	.01

The average parameter estimates of the allele frequencies and distinct disequilibrium coefficients with corresponding standard errors are also provided in Appendix A. Again, all parameters estimates appear unbiased with standard errors decreasing with the sample size.

Linkage disequialibria was tested with the same five χ^2 statistics as in the first simulation. Here we observe the three multiallelic marker test statistic, T_1 to be considerably more powerful than the other tests.

DOI: 10.2202/1544-6115.1528

Table 3: Measuring the power of test statistics T_1 through T_5.

$\boldsymbol{n = 30}$	T_1	T_2	T_3	T_4	T_5
$\Pr[T < .05]$	.579	.290	.363	.272	.224
$\Pr[T < .01]$	.262	.103	.175	.107	.073
$\Pr[T < .001]$	.070	.016	.047	.028	.007

$\boldsymbol{n = 100}$	T_1	T_2	T_3	T_4	T_5
$\Pr[T < .05]$	1.000	.743	.827	.705	.504
$\Pr[T < .01]$	.998	.513	.641	.474	.246
$\Pr[T < .001]$	.975	.239	.368	.214	.089
$\Pr[T < 10^{-6}]$	.513	.015	.036	.011	.001

$\boldsymbol{n = 200}$	T_1	T_2	T_3	T_4	T_5
$\Pr[T < .05]$	1.000	.978	.990	.958	.783
$\Pr[T < .01]$	1.000	.909	.950	.866	.580
$\Pr[T < .001]$	1.000	.688	.832	.665	.324
$\Pr[T < 10^{-6}]$	.999	.140	.273	.122	.027

For this particular simulation, the second best test is T_3, a test with two multiallelic markers. However, since all three pairs would be computed, some type of multiple testing correction should be invoked which would reduce the power of the two multiallelic marker test for linkage disequilibrium.

It should be pointed out that the simulations of the type-I errors indicated the test statistics had slightly varying nominal levels which can lead to more substantial differences in power. In Table 3, we compare the type-II errors across the different tests, but with the caveat that the tests varied in their nominal level. Even knowing variation is present in the nominal levels, it remains quite clear that T_1 has the best performance.

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; Zaykin *et al.* 2008). 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.* 1996; 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).

The past several decades shows a tremendous interest in developing analytical methods for estimating and testing linkage disequilibria (Weir and Cockerham 1978; Weir and Ott 1996; Excoffier 1995; Li *et al.* 2007; Zaykin *et al.* 2008; Li *et al.* 2009; Dupuis *et al.* 2007; Georges 2007; Kurbasic and Hossjer 2008). Most of these methods deal with linkage disequilibria between two or more markers each with two alternative alleles, and some of them are extended to model linkage disequilibria between two markers with multiple alleles (≥ 3). However, no method has been available yet to analyze linkage disequilibria of different orders among three different markers. This article for the first time presents a model for exploring the feasibility of linkage disequilibrium analysis with three multiallelic markers, and it provides a general methodology to extend multiallelic LD estimation to an arbitrary number of loci. Simulation studies showed that the new model has better power for disequilibrium detection and better precision for disequilibrium estimation compared with conventional two-marker analysis. We also found that linkage disequilibrium analysis of multiallelic markers by collapsing multiple alleles into two different categories may lose much information that is contained within multilocus multiallelic marker data. The rapid convergence of the EM algorithm easily allows hundreds of tests to be preformed within hours on an ordinary laptop. The new model is validated by a small data set collected for a plant outcrossing species, leading to the detection of significant linkage disequilibria among microsatellite markers.

With a considerable use of multiallelic markers, such as microsatellites, in outcrossing populations, three marker multiallic analysis is crucial to the accurate estimation and testing of linkage disequilibria. Recent development of array biotechnologies has made it possible to produce a massive amount of single nucleotide polymorphism (SNP) data, providing fuel for studying the pattern of genetic variation in the genome. Although SNPs are biallelic, they are often analyzed at the haplotype level. Haplotypes, i.e., combination of alleles at different markers on the same chromosome, are thought to explain genetic variation in complex traits and diseases (Liu *et al.* 2004; Wu *et al.* 2007; Rha *et al.* 2007). When multiple SNPs are modeled by haplotypes, haplotypes will function as if they are different alleles. Thus, our model proposed

DOI: 10.2202/1544-6115.1528

will find immediate application in studying the pattern of genetic structure and diversity in a natural population with increasing available SNP data.

APPENDIX A – Supplementary Tables from Simulation

Below are the average parameter estimates and standard errors of the allele frequencies and distinct disequilibrium coefficients under the null of no linkage disequilibrium. These estimates were computed with the EM algorithm provided in Section 2 of the main article.

Table 4: Allele frequency estimates and standard errors under equilibrium

$\boldsymbol{n = 30}$	p_1	p_2	q_1	q_2	q_3	r_1	r_2	r_3	r_4
true value	**.75**	**.25**	**.3**	**.3**	**.4**	**.5**	**.2**	**.1**	**.2**
avg estimate	.749	.251	.301	.300	.399	.500	.200	.100	.200
std err	.056	.056	.060	.060	.064	.064	.052	.039	.052

$\boldsymbol{n = 100}$	p_1	p_2	q_1	q_2	q_3	r_1	r_2	r_3	r_4
true value	**.75**	**.25**	**.3**	**.3**	**.4**	**.5**	**.2**	**.1**	**.2**
avg estimate	.750	.250	.300	.300	.400	.500	.200	.100	.200
std err	.030	.030	.032	.032	.035	.035	.028	.021	.028

$\boldsymbol{n = 200}$	p_1	p_2	q_1	q_2	q_3	r_1	r_2	r_3	r_4
true value	**.75**	**.25**	**.3**	**.3**	**.4**	**.5**	**.2**	**.1**	**.2**
avg estimate	.750	.250	.300	.300	.400	.500	.200	.100	.200
std err	.022	.022	.023	.023	.025	.025	.020	.015	.020

Table 5: Disequilibrium coefficient estimates and standard errors under equilibrium

$\boldsymbol{n = 30}$	$\mathcal{D}_{111}$	$\mathcal{D}_{112}$	$\mathcal{D}_{113}$	$\mathcal{D}_{114}$	$\mathcal{D}_{121}$	$\mathcal{D}_{122}$	$\mathcal{D}_{123}$	$\mathcal{D}_{124}$	$\mathcal{D}_{131}$
true value	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
avg estimate	.003	-.001	-.001	-.001	.001	-.001	-.001	-.001	.003
std err	.049	.034	.023	.035	.049	.034	.023	.034	.054

	$\mathcal{D}_{132}$	$\mathcal{D}_{133}$	$\mathcal{D}_{211}$	$\mathcal{D}_{212}$	$\mathcal{D}_{213}$	$\mathcal{D}_{221}$	$\mathcal{D}_{222}$	$\mathcal{D}_{223}$
true value	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
avg estimate	.000	-.001	-.002	.001	.001	-.002	.001	.001
std err	.038	.026	.034	.021	.015	.033	.022	.015

$\boldsymbol{n = 100}$	$\mathcal{D}_{111}$	$\mathcal{D}_{112}$	$\mathcal{D}_{113}$	$\mathcal{D}_{114}$	$\mathcal{D}_{121}$	$\mathcal{D}_{122}$	$\mathcal{D}_{123}$	$\mathcal{D}_{124}$	$\mathcal{D}_{131}$
true value	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
avg estimate	.001	.000	-.001	.000	.001	.000	-.001	.000	-.001
std err	.027	.020	.015	.020	.027	.020	.015	.020	.029

	$\mathcal{D}_{132}$	$\mathcal{D}_{133}$	$\mathcal{D}_{211}$	$\mathcal{D}_{212}$	$\mathcal{D}_{213}$	$\mathcal{D}_{221}$	$\mathcal{D}_{222}$	$\mathcal{D}_{223}$
true value	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
avg estimate	.000	.000	.000	.000	.000	-.001	.000	.000
std err	.022	.016	.021	.014	.009	.021	.014	.009

$\boldsymbol{n = 200}$	$\mathcal{D}_{111}$	$\mathcal{D}_{112}$	$\mathcal{D}_{113}$	$\mathcal{D}_{114}$	$\mathcal{D}_{121}$	$\mathcal{D}_{122}$	$\mathcal{D}_{123}$	$\mathcal{D}_{124}$	$\mathcal{D}_{131}$
true value	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
avg estimate	.000	.000	.000	.000	.000	.000	.000	.000	.000
std err	.019	.014	.011	.014	.018	.014	.010	.016	.019

	$\mathcal{D}_{132}$	$\mathcal{D}_{133}$	$\mathcal{D}_{211}$	$\mathcal{D}_{212}$	$\mathcal{D}_{213}$	$\mathcal{D}_{221}$	$\mathcal{D}_{222}$	$\mathcal{D}_{223}$
true value	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
avg estimate	.000	.000	.000	.000	.000	.000	.000	.000
std err	.015	.011	.015	.011	.007	.014	.010	.007

DOI: 10.2202/1544-6115.1528

Below are the average parameter estimates and standard errors of the allele frequencies and distinct disequilibrium coefficients simulated under the alternative hypothesis of linkage disequilibrium.

Table 6: Allele frequency estimates and standard errors under disequilibrium

$\boldsymbol{n = 30}$	p_1	p_2	q_1	q_2	q_3	r_1	r_2	r_3	r_4
true value	**.75**	**.25**	**.3**	**.3**	**.4**	**.5**	**.2**	**.1**	**.2**
avg estimate	.748	.252	.299	.300	.401	.498	.202	.100	.199
std err	.057	.057	.059	.058	.062	.066	.052	.039	.053

$\boldsymbol{n = 100}$	p_1	p_2	q_1	q_2	q_3	r_1	r_2	r_3	r_4
true value	**.75**	**.25**	**.3**	**.3**	**.4**	**.5**	**.2**	**.1**	**.2**
avg estimate	.752	.248	.298	.302	.400	.500	.200	.101	.200
std err	.030	.030	.032	.032	.031	.035	.035	.036	.027

$\boldsymbol{n = 200}$	p_1	p_2	q_1	q_2	q_3	r_1	r_2	r_3	r_4
true value	**.75**	**.25**	**.3**	**.3**	**.4**	**.5**	**.2**	**.1**	**.2**
avg estimate	.750	.250	.300	.300	.400	.501	.199	.100	.200
std err	.022	.022	.023	.023	.024	.025	.020	.015	.020

Table 7: Disequilibrium coefficient estimates and standard errors under disequilibrium

$\boldsymbol{n = 30}$	$\mathcal{D}_{111}$	$\mathcal{D}_{112}$	$\mathcal{D}_{113}$	$\mathcal{D}_{114}$	$\mathcal{D}_{121}$	$\mathcal{D}_{122}$	$\mathcal{D}_{123}$	$\mathcal{D}_{124}$	$\mathcal{D}_{131}$
true value	**.01**	**-.02**	**-.01**	**-.04**	**-.02**	**.03**	**0**	**0**	**.05**
avg estimate	.010	-.021	-.010	-.034	-.016	.027	.001	-.004	.050
std err	.044	.026	.018	.023	.046	.033	.024	.031	.050

	$\mathcal{D}_{132}$	$\mathcal{D}_{133}$	$\mathcal{D}_{211}$	$\mathcal{D}_{212}$	$\mathcal{D}_{213}$	$\mathcal{D}_{221}$	$\mathcal{D}_{222}$	$\mathcal{D}_{223}$
true value	**-.06**	**-.01**	**0**	**.04**	**0**	**0**	**-.01**	**.01**
avg estimate	-.054	.007	-.001	.040	.002	-.005	-.005	.008
std err	.024	.026	.032	.034	.016	.028	.020	.018

$\boldsymbol{n = 100}$	$\mathcal{D}_{111}$	$\mathcal{D}_{112}$	$\mathcal{D}_{113}$	$\mathcal{D}_{114}$	$\mathcal{D}_{121}$	$\mathcal{D}_{122}$	$\mathcal{D}_{123}$	$\mathcal{D}_{124}$	$\mathcal{D}_{131}$
true value	**.01**	**-.02**	**-.01**	**-.04**	**-.02**	**.03**	**0**	**0**	**.05**
avg estimate	.010	-.020	-.010	-.039	-.020	.030	.000	.000	.050
std err	.020	.013	.011	.011	.022	.016	.014	.016	.023

	$\mathcal{D}_{132}$	$\mathcal{D}_{133}$	$\mathcal{D}_{211}$	$\mathcal{D}_{212}$	$\mathcal{D}_{213}$	$\mathcal{D}_{221}$	$\mathcal{D}_{222}$	$\mathcal{D}_{223}$
true value	**-.06**	**-.01**	**0**	**.04**	**0**	**0**	**-.01**	**.01**
avg estimate	-.059	.010	.001	.040	.000	.000	-.009	.009
std err	.009	.014	.018	.018	.008	.015	.010	.011

$\boldsymbol{n = 200}$	$\mathcal{D}_{111}$	$\mathcal{D}_{112}$	$\mathcal{D}_{113}$	$\mathcal{D}_{114}$	$\mathcal{D}_{121}$	$\mathcal{D}_{122}$	$\mathcal{D}_{123}$	$\mathcal{D}_{124}$	$\mathcal{D}_{131}$
true value	**.01**	**-.02**	**-.01**	**-.04**	**-.02**	**.03**	**0**	**0**	**.05**
avg estimate	.009	-.020	-.010	-.040	-.019	.029	.000	.000	.050
std err	.014	.009	.007	.008	.015	.011	.009	.011	.016

	$\mathcal{D}_{132}$	$\mathcal{D}_{133}$	$\mathcal{D}_{211}$	$\mathcal{D}_{212}$	$\mathcal{D}_{213}$	$\mathcal{D}_{221}$	$\mathcal{D}_{222}$	$\mathcal{D}_{223}$
true value	**-.06**	**-.01**	**0**	**.04**	**0**	**0**	**-.01**	**.01**
avg estimate	-.060	.010	.001	.040	.000	.000	-.010	.010
std err	.006	.010	.012	.012	.006	.010	.007	.008

DOI: 10.2202/1544-6115.1528

APPENDIX B - *Lycoris longituba* Microsatellite Data

Table 8: Microsatellite data for *Lycoris longituba* used in the worked example.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}	M_{16}	M_{17}	M_{18}
1	BC	BB	..	..	..	CC	AA	CC	BB	AA	BB	..	CC	..	..	CD	..	AA
2	CC	AB	BC	..	AA	CC	BD	..	..	BC	AA	AA	AB	BD	BB	..	BC	AC
3	BC	..	BB	..	BC	CC	AA	BB	..	AA	BB	..	BC	CC	BC	AD	DD	BC
4	BC	BB	BB	BB	AA	BC	AA	AA	CC	AA	AB	BB	CC	BC	BB	BC	CC	CC
5	BC	BB	CC	AC	AB	BC	BB	BB	BB	BB	AB	CC	BC	CC	..	CC	AD	CC
6	BB	AB	BB	BB	AB	CC	BB	CC	CC	CC	BB	BC	BC	CD	AA	BC	AA	AA
7	BD	BB	BC	AC	BC	CC	BC	DD	BB	BC	AC	BC	BC	CC	..	BC	CC	AA
8	AB	AB	BB	AC	BB	BC	BB	BB	CC	BB	BB	BC	CC	CC	BC	BC	..	AA
9	AA	AB	BB	AC	BB	BC	BB	AA	BC	AA	AB	AC	BC	CC	BB	AD	AD	AA
10	BB	AB	AB	BD	AB	CC	AC	BC	BB	BB	CC	CC	CC	CC	CC	BC	AD	CC
11	AA	BB	AA	CC	BC	BC	AB	BD	BB	AA	AA	AC	BB	CC	..	BC	AB	CC
12	AB	AA	BB	AC	AB	CC	BC	AA	AB	BB	AB	BC	BC	CC	CC	BC	DD	AC
13	AB	AB	AB	AC	AB	BC	BC	BB	..	BB	AB	CC	..	CC	CC	CC	AD	..
14	AA	BB	AA	DD	AC	BC	BB	AC	BB	AA	AB	CC	BB	BC	BB	AA	BC	AC
15	AA	AB	AA	..	AB	BB	BB	DD	BB	BB	BB	CC	BB	AC	AB	AD	BB	AA
16	AB	BB	AA	CD	AA	BC	..	CC	BB	BB	..	CC	BB	CC	..	AC	AB	AB
17	AB	BB	AA	AC	BB	..	BB	AC	BC	BB	AB	CC	BC	CC	BB	AC	AB	AA
18	AC	AB	AB	AC	CC	BB	BC	DD	CC	AC	BB	BC	..	CC	AB	BC	CD	..
19	AA	AB	AA	AC	BB	BC	AA	BC	BC	BB	CC	CC	CC	CC	AB	BC	AC	AC
20	BB	BB	AA	AC	AB	BC	BB	BB	AC	BC	BB	..	BB	CC	..	BC	AA	AC
21	BB	AB	AA	AC	AC	AB	BB	CC	BC	AC	BB	BC	AB	CD	AB	..	BB	AB
22	BB	BB	AA	BC	AC	BC	BB	AA	BC	BB	AA	BC	BC	CC	AA	BC	AA	AB
23	AD	AB	AC	BC	BC	BB	AD	CC	CC	BB	AB	BC	BC	BC	BB	BC	BC	BB
24	AB	AB	AA	AB	AA	BB	AA	AA	BC	BC	CC	CC	BC	CC	BB	AC	AB	AA
25	AD	BB	AC	AB	BB	BB	AD	CC	AC	AB	BB	BB	CC	CC	BC	AD	AB	..
26	AC	BB	AC	BB	BB	BB	AC	DD	BC	BB	BB	BC	BB	CC	AB	AA	AD	AA
27	BC	AB	..	CC	BB	AB	BB	AC	BC	BB	..	BC	CC	CC	AB	BC	AD	AA
28	BC	BB	BB	CD	AA	AB	BB	AA	CC	BC	BB	BC	CC	CC	AB	BC	AD	AB
29	AB	AB	AA	BC	BB	AB	AA	BB	BB	BB	AA	CC	CC	CC	BC	AD	AB	AB
30	CC	BB	BB	BC	AA	AB	BB	AA	CC	BB	BB	CC	BB	CC	BC	..	BB	AA
31	BC	BB	BB	..	BB	BB	BB	BB	CC	..	BB	BC	CC	CC	BC	BC	AD	AC
32	BD	..	BC	BC	AA	BB	BD	AB	BB	..	AA	CC	..	CC	BB	AD	AB	AA

APPENDIX C – Trigenic Disequilibrium Estimation Details

The general haplotype model for the $m_1 m_2 m_3$ haplotypes is re-written as

$$p_{ijk} = p_i q_j r_k + \tilde{D}_{ijk} \tag{3}$$

(for $i = 1, \ldots, m_1; j = 1, \ldots, m_2; k = 1, \ldots, m_3$) where $\tilde{D}_{ijk}$ is a reparameterization of the digenic and trigenic LD coefficients. In order to reduce the linkage disequlibrium model to an identifiable model, the reparameterized trigenic linkage disequilibrium coefficients, $\tilde{D}_{ijk}$, are defined in terms of the variables $\mathcal{D}_{ijk}$. Specifically, we define $\tilde{D}_{ijk}$ as a function of $\mathcal{D}_{ijk}$ with the following recursive expression.

$$\tilde{D}_{ijk} = \begin{cases} \mathcal{D}_{ijk}, & (i,j,k) \in S \\ -\sum_{j'+k'<m_2+m_3} \tilde{D}_{ij'k'}, & i < m_1 \;\&\; j = m_2 \;\&\; k = m_3 \\ -\sum_{i'+k'<m_1+m_3} \tilde{D}_{i'jk'}, & i = m_1 \;\&\; j < m_2 \;\&\; k = m_3 \\ -\sum_{i'+j'<m_1+m_2} \tilde{D}_{i'j'k}, & i = m_1 \;\&\; j = m_2 \;\&\; k < m_3 \\ -\sum_{(i',j',k')\neq(m_1,m_2,m_3)} \tilde{D}_{i'j'k'}, & i = m_1 \;\&\; j = m_2 \;\&\; k = m_3. \end{cases} \tag{4}$$

where S is the triple of indices where at most one index meets its maximum, i.e.

$$\begin{aligned} S :=& \{(i,j,k) \in \mathbb{Z}_+^3 | (i < m_1 \;\&\; j < m_2) \text{ or} \\ & (i < m_1 \;\&\; k < m_3) \text{ or } (j < m_2 \;\&\; k < m_3)\}. \end{aligned}$$

The motivation for the above parameterization is presented. The recursion relations can be inferred from properties such as the following

$$p_i = \sum_{j,k} p_{i,j,k} = \sum_{j,k} p_i q_j r_k + \sum_{j,k} \tilde{D}_{ijk} = p_i + \sum_{j,k} \tilde{D}_{ijk}$$

DOI: 10.2202/1544-6115.1528

which leads to the recursion found in the second case, i.e. $\tilde{D}_{im_2m_3}$ for $i < m_1$.

The number of distinct trigenic disequilibrium coefficients equals the cardinality of S which can be calculated as

$$\begin{aligned}|S| &= (m_1-1)(m_2-1)(m_3-1) \\ &\quad + (m_1-1)(m_2-1) + (m_2-1)(m_3-1) + (m_1-1)(m_3-1) \\ &= m_1m_2m_3 - m_1 - m_2 - m_3 + 2\end{aligned}$$

where the first summand of $|S|$ comes from the number of distinct tuples with indices all smaller than their maximum and the other summands come from tuples where exactly one index is at its maximum. An alternative calculation notes that the $m_1m_2m_3 - 1$ degrees of freedom are taken up by $(m_1 - 1) + (m_2 - 1) + (m_3 - 1)$ allele probabilities for the three markers (the number of independent p's, q's, and r's) and by the distinct disequilibrium parameters, again giving the number of distinct coefficients to be

$$\begin{aligned}&\underbrace{m_1m_2m_3 - 1}_{\text{degrees of freedom}} - \underbrace{((m_1-1) + (m_2-1) + (m_3-1))}_{\text{number of independent } p\text{'s, } q\text{'s, and } r\text{'s}} \\ &\quad = \underbrace{m_1m_2m_3 - m_1 - m_2 - m_3 + 2}_{\text{number of distinct } D\text{'s}}.\end{aligned}$$

The EM algorithm is utilized to derive the maximum likelihood estimates (MLEs) of the haplotype frequencies as follows:

Step I. Start with an initial estimate of p_{ijk}; for instance, $\hat{p}_{ijk} = \frac{1}{m_1m_2m_3}$ for all i, j, and k.

Step II. Compute $\hat{\phi}(i,j,k) = \hat{\phi}_{i_1 i_2 j_1 j_2 k_1 k_2}(i,j,k)$ where $\hat{\phi}(i,j,k) =$

$$\begin{cases} 2\hat{P}_{i_1 i_2 j_1 j_2 k_1 k_2} & i_1 = i_2 = i \;\&\; j_1 = j_2 = j \;\&\; k_1 = k_2 = k \\ \hline \hat{P}_{i_1 i_2 j_1 j_2 k_1 k_2} & \big(\text{exactly two are true: } i_1 = i_2,\, j_1 = j_2,\, k_1 = k_2\big) \;\& \\ & \quad (i_1 = i \text{ or } i_2 = i) \;\&\; (j_1 = j \text{ or } j_2 = j) \;\&\; (k_1 = k \text{ or } k_2 = k) \\ \hline \hat{p}_{ij_1k_1}\hat{p}_{ij_2k_2} & i_1 = i_2 = i \;\&\; j_1 \neq j_2 \;\&\; k_1 \neq k_2 \;\& \\ & \quad \big((j_1 = j \;\&\; k_1 = k) \text{ or } (j_2 = j \;\&\; k_2 = k)\big) \\ \hline \hat{p}_{ij_1k_2}\hat{p}_{ij_2k_1} & i_1 = i_2 = i \;\&\; j_1 \neq j_2 \;\&\; k_1 \neq k_2 \;\& \\ & \quad \big((j_1 = j \;\&\; k_2 = k) \text{ or } (j_2 = j \;\&\; k_1 = k)\big) \\ \hline \hat{p}_{i_1jk_1}\hat{p}_{i_2jk_2} & i_1 \neq i_2 \;\&\; j_1 = j_2 = j \;\&\; k_1 \neq k_2 \;\& \\ & \quad \big((i_1 = i \;\&\; k_1 = k) \text{ or } (i_2 = i \;\&\; k_2 = k)\big) \\ \hline \hat{p}_{i_1jk_2}\hat{p}_{i_2jk_1} & i_1 \neq i_2 \;\&\; j_1 = j_2 = j \;\&\; k_1 \neq k_2 \;\& \\ & \quad \big((i_1 = i \;\&\; k_2 = k) \text{ or } (i_2 = i \;\&\; k_1 = k)\big) \\ \hline \hat{p}_{i_1j_1k}\hat{p}_{i_2j_2k} & i_1 \neq i_2 \;\&\; j_1 \neq j_2 \;\&\; k_1 = k_2 = k \;\& \\ & \quad \big((i_1 = i \;\&\; j_1 = j) \text{ or } (i_2 = i \;\&\; j_2 = j)\big) \\ \hline \hat{p}_{i_1j_1k}\hat{p}_{i_2j_1k} & i_1 \neq i_2 \;\&\; j_1 \neq j_2 \;\&\; k_1 = k_2 = k \;\& \\ & \quad \big((i_1 = i \;\&\; j_2 = j) \text{ or } (i_2 = i \;\&\; j_1 = j)\big) \\ \hline \hat{p}_{i_1j_1k_1}\hat{p}_{i_2j_2k_2} & i_1 \neq i_2 \;\&\; j_1 \neq j_2 \;\&\; k_1 \neq k_2 \;\& \\ & \quad \big((i_1 = i \;\&\; j_1 = j \;\&\; k_1 = k) \text{ or } (i_2 = i \;\&\; j_2 = j \;\&\; k_2 = k)\big) \\ \hline \hat{p}_{i_1j_1k_2}\hat{p}_{i_2j_2k_1} & i_1 \neq i_2 \;\&\; j_1 \neq j_2 \;\&\; k_1 \neq k_2 \;\& \\ & \quad \big((i_1 = i \;\&\; j_1 = j \;\&\; k_2 = k) \text{ or } (i_2 = i \;\&\; j_2 = j \;\&\; k_1 = k)\big) \\ \hline \hat{p}_{i_1j_2k_1}\hat{p}_{i_2j_1k_2} & i_1 \neq i_2 \;\&\; j_1 \neq j_2 \;\&\; k_1 \neq k_2 \;\& \\ & \quad \big((i_1 = i \;\&\; j_2 = j \;\&\; k_1 = k) \text{ or } (i_2 = i \;\&\; j_1 = j \;\&\; k_2 = k)\big) \\ \hline \hat{p}_{i_1j_2k_2}\hat{p}_{i_2j_1k_1} & i_1 \neq i_2 \;\&\; j_1 \neq j_2 \;\&\; k_1 \neq k_2 \;\& \\ & \quad \big((i_1 = i \;\&\; j_2 = j \;\&\; k_2 = k) \text{ or } (i_2 = i \;\&\; j_1 = j \;\&\; k_1 = k)\big) \\ \hline 0 & \text{otherwise} \end{cases}$$

where $\hat{P}_{i_1 i_2 j_1 j_2 k_1 k_2}$ be the fraction of observed multilocus genotypes of type $i_1 i_2 j_1 j_2 k_1 k_2$.

Step III. Update $\hat{p}_{ijk}$ from the equation

$$\hat{p}_{ijk} = \frac{1}{2n} \sum \frac{\hat{\phi}(i,j,k)}{\hat{P}_{i_1 i_2 j_1 j_2 k_1 k_2}} n_{i_1 i_2 j_1 j_2 k_1 k_2}$$

where the sum is over all valid genotypes $i_1 i_2 j_1 j_2 k_1 k_2$ and n is the number of individuals.

Step IV. Repeat steps II and III until the parameters have converged.

Convergence of the EM algorithm above leads to estimates of the haplotype probabilities. In order to estimate the linkage disequilibria coefficients, estimates of the allele frequencies are needed. The allele frequencies can then

DOI: 10.2202/1544-6115.1528

be estimated by

$$\begin{aligned}
\hat{p}_i &= \frac{1}{2n} \quad n_{ii\cdot\cdot\cdot\cdot} + \sum_{i_1=1}^{m_1} \sum_{i_2=i_1}^{m_1} 1_{(i_1=i \text{ or } i_2=i)} n_{i_1 i_2 \cdot\cdot\cdot\cdot} \Bigg) \\
\hat{q}_j &= \frac{1}{2n} \quad n_{\cdot\cdot jj\cdot\cdot} + \sum_{j_1=1}^{m_2} \sum_{j_2=j_1}^{m_2} 1_{(j_1=j \text{ or } j_2=j)} n_{\cdot\cdot j_1 j_2 \cdot\cdot} \Bigg) \\
\hat{r}_k &= \frac{1}{2n} \quad n_{\cdot\cdot\cdot\cdot kk} + \sum_{k_1=1}^{m_3} \sum_{k_2=k_1}^{m_3} 1_{(k_1=k \text{ or } k_2=k)} n_{\cdot\cdot\cdot\cdot k_1 k_2} \Bigg)
\end{aligned} \tag{5}$$

where the usual dot notation represents the sum over the given variable; e.g. $n_{i_1 i_2 \cdot\cdot\cdot\cdot}$ is the sum over j_1, j_2, k_1, and k_2. Therefore estimates of the linkage disequilibrium coefficients can be estimated as

$$\widehat{\tilde{D}}_{ijk} = \hat{p}_{ijk} - \hat{p}_i \hat{q}_j \hat{r}_k.$$

In the case of two multiallelic markers, the haplotype model, similar to (3) is given by $p_{ij} = p_i q_j + D_{ij}$ where D_{ij} is the two-locus linkage disequilibrium that can be estimated through similar means as in Kalinowski and Hedrick (2001). With estimates of the digenic linkage disequilibria, $\hat{D}_{ij}$, the estimates of the trigenic linkage disequilibrium coefficients, $\hat{D}_{ijk}$, are then easily computed from (3) as

$$\hat{D}_{ijk} = (-1)^{i+j+k} \widehat{\tilde{D}}_{ijk} + (-1)^{k+1} \hat{D}_{ij} + (-1)^{j+1} \hat{D}_{ik} + (-1)^{i+1} \hat{D}_{jk}.$$

LITERATURE CITED

Ardlie K, Kruglyak L, Seielstad M, *et al.* (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* **3**: 299–309.

Bennett J (1954). On the theory of random mating. *Ann Eugen* **18**: 311–7.

Dausset J, Legrand L, Lepage V, Contu L, Marcelli-Barge A, Wildloecher I, *et al.* (1978). A haplotype study of hla complex with special reference to the hla-dr series and to bf. c2 and glyoxalase i polymorphisms. *Tissue Antigens* **12**: 297–307.

Dawson E, Abecasis G, Bumpstead S, Chen Y, Hunt S, Beare D, *et al.* (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**: 544–548.

Dupuis J, Siegmund D, Yakir B (2007). A unified framework for linkage and association analysis of quantitative traits. *Proceedings of the National Academy of Sciences* **104**: 20210–20215.

Excoffier L (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**: 921–927.

Excoffier L, Slatkin M (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular biology and evolution* **12**: 921.

Gabriel S, Schaffner S, Nguyen H, Moore J, Roy J, Blumenstiel B, *et al.* (2002). The structure of haplotype blocks in the human genome. *Science* **296**: 2225–2229.

Geiringer H (1944). On the probability theory of linkage in mendelian heredity. *The Annals of Mathematical Statistics* pp. 25–57.

Georges M (2007). Mapping, fine mapping, and molecular dissection of quantitative trait loci in domestic animals. *Annual Review of Genomics and Human Genetics* **8**: 131–162.

Gorelick R, Laubichler M (2004). Decomposing multilocus linkage disequilibrium. *Genetics* **166**: 1581.

Guo S, Thompson E (1992). Performing the exact test of hardy-weinberg proportion for multiple alleles. *Biometrics* **48**: 361–372.

He Q, Shen Y, Chen Y, Zhou Y, Berg A, Wu R, *et al.* (2009). Development of 16 polymorphic simple sequence repeat markers for lycoris longituba from expressed sequence tags. *Molecular Ecology Resources* **9**: 278–280.

Hernandez J, Weir B (1989). A disequilibrium coefficient approach to hardy-weinberg testing. *Biometrics* **45**: 53–70.

Hill W (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity* **33**: 229–239.

Hössjer O, Hartman L, Humphreys K (2009). Ancestral Recombination Graphs under Non-Random Ascertainment, with Applications to Gene Mapping. *Statistical Applications in Genetics and Molecular Biology* **8**: 35.

Kalinowski S, Hedrick P (2001). Estimation of linkage disequilibrium for loci with multiple alleles: basic approach and an application using data from bighorn sheep. *Heredity* **87**: 698–708.

Kim Y, Feng S, Zeng Z (2008). Measuring and partitioning the high-order linkage disequilibrium by multiple order markov chains. *Genetic Epidemiology* **32**.

Kurbasic A, Hossjer O (2008). A general method for linkage disequilibrium correction for multipoint linkage and association. *Genetic Epidemiology* **32**: 647–657.

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

Li J, Li Q, Hou W, Han K, Li Y, Wu S, *et al.* (2009). An algorithmic model for

constructing a linkage and linkage disequilibrium map in outcrossing plant populations. *Genetics Research* **91**: 9–21.

Li Q, Wu R (2009). A multilocus model for constructing a linkage disequilibrium map in human populations. *Statistical Applications in Genetics and Molecular Biology* **8**: Iss. 1, Article 18. DOI: 10.2202/1544-6115.1419.

Li Y, Li Y, Wu S, Han K, Wang Z, Hou W, *et al.* (2007). Estimation of multilocus linkage disequilibria in diploid populations with dominant markers. *Genetics* **176**: 1811–1821.

Liu T, Johnson J, Casella G, Wu R (2004). Sequencing complex diseases with hapmap. *Genetics* **168**: 503–511.

Long J, Williams R, Urbanek M (1995). An em algorithm and testing strategy for multiple-locus haplotypes. *American Journal of Human Genetics* **56**: 799–810.

Louis E, Dempster E (1987). An exact test for hardy-weinberg and multiple alleles. *Biometrics* **43**: 805–811.

Lynch M, Walsh B (1998). *Genetics and analysis of quantitative traits.* Sinauer Associates, Inc.

Miller A, Schaal B (2006). Domestication and the distribution of genetic variation in wild and cultivated populations of the mesoamerican fruit tree spondias purpurea l.(anacardiaceae). *Molecular Ecology* **15**: 1467–1480.

Mohlke K, Lange E, Valle T, Ghosh S, Magnuson V, Silander K, *et al.* (2001). Linkage disequilibrium between microsatellite markers extends beyond 1 cm on chromosome 20 in finns. *Genome Research* **11**: 1221–1226.

Rafalski A, Morgante M (2004). Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. *Trends in Genetics* **20**: 103–111.

Rha S, Jeung H, Choi Y, Yang W, Yoo J, Kim B, *et al.* (2007). An association between rrm1 haplotype and gemcitabine-induced neutropenia in breast cancer patients. *The Oncologist* **12**: 622–630.

Robinson W, Asmussen M, Thomson G (1991). Three-locus systems impose additional constraints on pairwise disequilibria. *Genetics* **129**: 925–930.

Slatkin M (2008). Linkage disequilibriumunderstanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* **9**: 477–486.

Slatkin M, Excoffier L (1996). Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. *Heredity* **76**: 377–383.

Stephens J, Schneider J, Tanguay D, Choi J, Acharya T, Stanley S, *et al.* (2001). Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**: 489–493.

Tishkoff S, Dietzsch E, Speed W, Pakstis A, Kidd J, Cheung K, *et al.* (1996). Global patterns of linkage disequilibrium at the cd4 locus and modern hu-

man origins. *Science* **271**: 1380–1387.

Weir B (1979). Inferences about linkage disequilibrium. *Biometrics* **35**: 235–254.

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

Weir B, Ott J (1996). *Genetic data analysis II*. Sinauer Associates Sunderland, MA.

Wu S, Yang J, Wang C, Wu R (2007). A general quantitative genetic model for haplotyping a complex trait in humans. *Current Genomics* **8**: 343–350.

Zaykin D, Pudovkin A, Weir B (2008). Correlation-based inference for linkage disequilibrium with multiple alleles. *Genetics* **180**: 533–545.

Zhivotovsky L (1999). Estimating population structure in diploids with multilocus dominant dna markers. *Molecular Ecology* **8**: 907–913.

DOI: 10.2202/1544-6115.1528