frontiers in GENETICS

METHODS ARTICLE
published: 02 July 2012
doi: 10.3389/fgene.2012.00078

Model and algorithm for linkage disequilibrium analysis in a non-equilibrium population

Jingyuan Liu*[1]*, Zhong Wang*[2]*, Yaqun Wang*[1]*, Runze Li*[1,2] *and Rongling Wu*[1,2] **

[1] Department of Statistics, The Pennsylvania State University, State College, PA, USA
[2] Division of Biostatistics and Bioinformatics, Pennsylvania State University, Hershey, PA, USA

Edited by:
Christina Kitchen, University of California Los Angeles, USA

Reviewed by:
Dalin Li, Cedars-Sinai Medical Center, USA
Christina Kitchen, University of California Los Angeles, USA

****Correspondence:***
Rongling Wu, Division of Biostatistics and Bioinformatics, Pennsylvania State University, Hershey, 17033 PA, USA.
e-mail: rwu@hes.hmc.psu.edu

The multilocus analysis of polymorphisms has emerged as a vital ingredient of population genetics and evolutionary biology. A fundamental assumption used for existing multilocus analysis approaches is Hardy–Weinberg equilibrium at which maternally- and paternally-derived gametes unite randomly during fertilization. Given the fact that natural populations are rarely panmictic, these approaches will have a significant limitation for practical use. We present a robust model for multilocus linkage disequilibrium analysis which does not rely on the assumption of random mating. This new disequilibrium model capitalizes on Weir's definition of zygotic disequilibria and is based on an open-pollinated design in which multiple maternal individuals and their half-sib families are sampled from a natural population. This design captures two levels of associations: one is at the upper level that describes the pattern of cosegregation between different loci in the parental population and the other is at the lower level that specifies the extent of co-transmission of non-alleles at different loci from parents to their offspring. An MCMC method was implemented to estimate genetic parameters that define these associations. Simulation studies were used to validate the statistical behavior of the new model.

Keywords: gametic linkage disequilibrium, zygotic linkage disequilibrium, Hardy–Weinberg equilibrium, non-equilibrium population, molecular marker

INTRODUCTION

Linkage disequilibria have been used as a fundamental concept for studying the pattern of genetic diversity in a natural population (Lewontin, 1964, 1988; Hedrick, 1987; Weir, 1996; Lou et al., 2003; Li et al., 2007; Slatkin, 2008) as well as for fine-mapping the genetic architecture of complex traits (Kruglyak, 1999). The traditional definition of linkage disequilibrium is the non-random association of alleles at different loci within the same gametes. The theoretical basis of estimating gametic linkage disequilibria is founded on the assumption that the population under study is at Hardy–Weinberg equilibrium (HWE), in which individuals are randomly mating to produce next generations. With the HWE assumption, genotype frequencies in the population are expressed as the product of gamete frequencies, a key expression to estimate and test the gametic linkage disequilibrium. Recently, Wu et al. (2010) have showed that, when multiple loci are considered simultaneously, this expression may not be necessarily true even if these loci are individually at HWE.

More importantly, for a given population, the assumption of random mating may be violated by many evolutionary forces such as selection, mutation, genetic drift, and population structure. For a non-equilibrium population at Hardy–Weinberg disequilibrium (HWD), zygotic disequilibria that have power to characterize non-random associations at both gametic and zygotic levels (Weir, 1996; Yang, 2000, 2002) may be more relevant. Earlier studies have documented possible genetic and evolutionary causes for zygotic associations in a non-equilibrium population (Haldane, 1949; Bennett and Binet, 1956; Charlesworth, 1991; Barton and Gale, 1993). Weir (1996) documented five different types of disequilibria simultaneously which are (1) Hardy–Weinberg disequilibria at each locus, (2) gametic disequilibrium (including two alleles in the same gamete, each from a different locus), (3) non-gametic disequilibrium (including two alleles in different gametes, each from a different locus), (4) trigenic disequilibrium (including a zygote at one locus and an allele at the other), and (5) quadrigenic disequilibrium (including two zygotes each from a different locus). Because it is impossible to estimate all the five disequilibrium parameters due to inadequate degrees of freedom, Weir (1996) collapsed gametic and non-gametic disequilibria to estimate a so-called composite gametic disequilibrium. More recently, Liu et al. (2006) used Weir's approach to estimate zygotic disequilibria in a canine population and gain a better insight into the structure and organization of the canine genome.

For a traditional population genetic strategy that samples unrelated individuals at random from a natural population, Weir's approach cannot separate the gametic and non-gametic disequilibria. This is because it provides insufficient information to distinguish two diplotypes of a double heterozygote which have the same genotype. In this article, we present a disequilibrium model for estimating the relative proportions of these two diplotypes for the double heterozygote and, therefore, making a distinction between the disequilibria occurring within and between gametes. The new model is based on an open-pollinated (OP) plant design that contains a set of randomly selected maternal plants from a natural population and multiple progeny from each maternal plant. The OP design was first proposed by Wu and Zeng (2001) to simultaneously estimate the linkage and (gametic) linkage disequilibrium between different markers. Li et al. (2009) further used this design to estimate the mating behavior of an outcrossing plant species, thus providing a comprehensive means for studying the genetic structure and diversity of natural populations. Hou et al. (2009) extended Li et al.'s (2009) work to jointly model the linkage, linkage disequilibrium, and genetic interference of genes using multilocus data. Here, we use the OP design to relax the HWE assumption which is needed for traditional multilocus analysis and modeling. While most of the previous work was implemented with the EM algorithm, we develop a Markov chain Monte Carlo (MCMC) procedure to estimate the parameters that define zygotic disequilibria. We conduct computer simulation to test the statistical properties of the model and validate its use in practice.

ZYGOTIC DISEQUILIBRIUM

GAMETE AND NON-GAMETE FREQUENCIES

Different types of zygotic disequilibria were defined in Weir (1996). As a follow-up of Liu et al.'s (2006) work, we will use the same notation and procedure for dissecting genotypic frequencies into disequilibrium components. Suppose we have two SNP markers **A** (with two alleles *A* and *a*) and **B** (with two alleles *B* and *b*). The frequencies of the corresponding alleles are denoted as p_A and p_a $(p_A + p_a = 1)$ as well as p_B and p_b $(p_B + p_b = 1)$, respectively. There are three distinguishable genotypes at each marker, i.e., *AA*, *Aa*, and *aa* for **A** and *BB*, *Bb*, and *bb* for **B**, whose genotype frequencies are denoted as *P* subscribed by the genotype notation. The two markers form 9 distinguishable genotypes, although there are 10 genotypic configurations or diplotypes. One genotype, *AaBb*, has two possible genotypic configurations ${}^{B}_{A}\left|{}^{b}_{a}\right|$ and ${}^{b}_{A}\left|{}^{B}_{a}\right|$, which are genotypically seen as the same.

Table 1 tabulates 9 genotype frequencies and 10 diplotype frequencies denoted by *P*, subscripted and superscripted by the genotype or diplotype notation. Using two-marker diplotype frequencies, we can estimate one-marker genotype frequencies by

$$\begin{aligned}
P_{AA} &= P^{BB}_{AA} + P^{Bb}_{AA} + P^{bb}_{AA} \\
P_{Aa} &= P^{BB}_{Aa} + P^{Bb}_{Aa} + P^{bB}_{Aa} + P^{bb}_{Aa} \\
P_{aa} &= P^{BB}_{aa} + P^{Bb}_{aa} + P^{bb}_{aa}
\end{aligned} \tag{1}$$

for marker **A**,

$$\begin{aligned}
P_{BB} &= P^{BB}_{AA} + P^{BB}_{Aa} + P^{BB}_{aa} \\
P_{Bb} &= P^{Bb}_{AA} + P^{Bb}_{Aa} + P^{bB}_{Aa} + P^{Bb}_{aa} \\
P_{bb} &= P^{bb}_{AA} + P^{bb}_{Aa} + P^{bb}_{aa}
\end{aligned} \tag{2}$$

for marker **B**, and further estimate the allele frequencies by

$$\begin{aligned}
p_A &= P_{AA} + \frac{1}{2}P_{Aa} \\
p_a &= P_{aa} + \frac{1}{2}P_{Aa} \\
p_B &= P_{BB} + \frac{1}{2}P_{Bb} \\
p_b &= P_{bb} + \frac{1}{2}P_{Bb}.
\end{aligned} \tag{3}$$

The two markers form four gametes, *AB*, *Ab*, *aB*, and *ab*, whose frequencies are denoted as *p* subscribed by the gamete notation. They are derived from diplotype frequencies by

$$\begin{aligned}
p_{AB} &= P^{BB}_{AA} + \frac{1}{2}\left(P^{Bb}_{AA} + P^{BB}_{Aa} + P^{Bb}_{Aa}\right) \\
p_{Ab} &= P^{bb}_{AA} + \frac{1}{2}\left(P^{Bb}_{AA} + P^{bb}_{Aa} + P^{bB}_{Aa}\right) \\
p_{aB} &= P^{BB}_{aa} + \frac{1}{2}\left(P^{BB}_{Aa} + P^{Bb}_{aa} + P^{bB}_{Aa}\right) \\
p_{ab} &= P^{bb}_{aa} + \frac{1}{2}\left(P^{bb}_{Aa} + P^{Bb}_{aa} + P^{Bb}_{Aa}\right).
\end{aligned} \tag{4}$$

Similarly, the frequencies of non-alleles from different gametes are derived as

$$\begin{aligned}
p_{A/B} &= P^{BB}_{AA} + \frac{1}{2}\left(P^{Bb}_{AA} + P^{BB}_{Aa} + P^{bB}_{Aa}\right) \\
p_{A/b} &= P^{bb}_{AA} + \frac{1}{2}\left(P^{Bb}_{AA} + P^{bb}_{Aa} + P^{Bb}_{Aa}\right) \\
p_{a/B} &= P^{BB}_{aa} + \frac{1}{2}\left(P^{BB}_{Aa} + P^{Bb}_{aa} + P^{Bb}_{Aa}\right) \\
p_{a/b} &= P^{bb}_{aa} + \frac{1}{2}\left(P^{bb}_{Aa} + P^{Bb}_{aa} + P^{bB}_{Aa}\right).
\end{aligned} \tag{5}$$

In Eqs 4 and 5, diplotype frequencies P^{Bb}_{Aa} and P^{bB}_{Aa} are mixed as a genotype frequency P_{AaBb}.

The frequencies of triple alleles from different markers are derived as

$$\begin{aligned}
p^{B}_{AA} &= P^{BB}_{AA} + \frac{1}{2}P^{Bb}_{AA}, & p^{b}_{AA} &= P^{bb}_{AA} + \frac{1}{2}P^{Bb}_{AA} \\
p^{B}_{Aa} &= P^{BB}_{Aa} + \frac{1}{2}\left(P^{Bb}_{Aa} + P^{bB}_{Aa}\right), & p^{b}_{Aa} &= P^{bb}_{Aa} + \frac{1}{2}\left(P^{Bb}_{Aa} + P^{bB}_{Aa}\right) \\
p^{B}_{aa} &= P^{BB}_{aa} + \frac{1}{2}P^{Bb}_{aa}, & p^{b}_{aa} &= P^{bb}_{aa} + \frac{1}{2}P^{Bb}_{aa} \\
p^{BB}_{A} &= P^{BB}_{AA} + \frac{1}{2}P^{BB}_{Aa}, & p^{BB}_{a} &= P^{BB}_{aa} + \frac{1}{2}P^{BB}_{Aa} \\
p^{Bb}_{A} &= P^{Bb}_{AA} + \frac{1}{2}\left(P^{Bb}_{Aa} + P^{bB}_{Aa}\right), & p^{Bb}_{a} &= P^{Bb}_{aa} + \frac{1}{2}\left(P^{Bb}_{Aa} + P^{bB}_{Aa}\right) \\
p^{bb}_{A} &= P^{bb}_{AA} + \frac{1}{2}P^{bb}_{Aa}, & p^{bb}_{a} &= P^{bb}_{aa} + \frac{1}{2}P^{bb}_{Aa}.
\end{aligned} \tag{6}$$

As shown above, all the genotype, allele, gamete, and non-gamete frequencies are uniquely determined by the diplotype frequencies. According to **Table 1**, the frequencies of all the diplotypes, except for *AB*|*ab* and *Ab*|*aB*, can be estimated from the observed data, since they are all one to one corresponding to genotypes which can be directly estimated from data.

A COMPLETE SET OF DISEQUILIBRIA

Zygotic disequilibrium is defined as the deviation of two-locus genotype frequencies from products of single-locus genotype frequencies and, thus, is composed of all non-allelic genic disequilibria at the two loci (Weir, 1996). For a population at HWD, the desirable property of an equilibrium population will not occur,

Table 1 | Frequencies and numbers of observations of marker genotypes.

Marker	**Marker B**			
A	**BB**	**Bb**	**bb**	**Total**
AA	$P_{AA}^{BB}(N_1)$	$P_{AA}^{Bb}(N_2)$	$P_{AA}^{bb}(N_3)$	$p_{AA}\ (N_1+N_2+N_3)$
Aa	$P_{Aa}^{BB}(N_4)$	$P_{Aa}^{Bb}+P_{Aa}^{bB}=P_{AaBb}(N_5)$	$P_{Aa}^{bb}(N_6)$	$P_{Aa}\ (N_4+N_5+N_6)$
aa	$P_{aa}^{BB}(N_7)$	$P_{aa}^{Bb}(N_8)$	$P_{aa}^{bb}(N_9)$	$P_{aa}\ (N_7+N_8+N_9)$
	P_{BB}	P_{Bb}	P_{bb}	*1*
Total	$\mathrm{n}_1+N_4+N_7$	$N_2+N_5+N_8$	$N_3+N_6+N_9$	*N.*

(1) Genotype AaBb contains two different configurations or diplotypes AB|ab and Ab|aB; (2) N$_i$'s in the parentheses are the numbers of corresponding observed genotypes.

such as independence of different alleles at the same locus (Lynch and Walsh, 1998). The HWD attempts to test for two alleles at the same locus, but on different gametes, whereas (gametic) linkage disequilibrium describes two alleles on the same gametes, but at different loci. For a zygotic disequilibrium, however, there is a third test, i.e., two alleles on different gametes and at different loci.

Since the population is not in HWE, two alleles at each marker are not independent, with the coefficients of HWD defined as

$$\begin{aligned} D_A &= P_{AA} - p_A^2 \\ &= -\frac{1}{2}P_{Aa} + p_A p_a \\ &= P_{aa} - p_a^2 \end{aligned} \tag{7}$$

for marker **A** and

$$\begin{aligned} D_B &= P_{BB} - p_B^2 \\ &= -\frac{1}{2}P_{Bb} + p_B p_b \\ &= P_{bb} - p_b^2 \end{aligned} \tag{8}$$

for marker **B**, respectively. The coefficient of digenic gametic linkage disequilibrium between the two markers is defined as

$$\begin{aligned} D_{ab} &= p_{AB} - p_A p_B \\ &= -p_{Ab} + p_A p_b \\ &= -p_{aB} + p_a p_B \\ &= p_{ab} - p_a p_b. \end{aligned} \tag{9}$$

For the non-equilibrium population, digenic linkage disequilibrium that occurs between non-alleles at different gametes is defined as

$$\begin{aligned} D_{a/b} &= p_{A/B} - p_A p_B \\ &= -p_{A/b} + p_A p_b \\ &= -p_{a/B} + p_a p_B \\ &= p_{a/b} - p_a p_b. \end{aligned} \tag{10}$$

The trigenic disequilibria between two alleles from marker **A** and one allele from marker **B** is defined as

$$\begin{aligned} D_{Ab} &= p_{AA}^{B} - p_A D_{ab} - p_A D_{a/b} - p_B D_A - p_A^2 p_B \\ &= -p_{AA}^{b} - p_A D_{ab} - p_A D_{a/b} + p_b D_A + p_A^2 p_b \\ &= -\frac{1}{2}p_{Aa}^{B} - \frac{1}{2}(p_A - p_a)D_{ab} - \frac{1}{2}(p_A - p_a)D_{a/b} \\ &\quad - p_B D_A + p_A p_a p_B \\ &= p_{aa}^{B} + p_a D_{ab} + p_a D_{a/b} - p_B D_A - p_a^2 p_B \\ &= -p_{aa}^{b} + p_a D_{ab} + p_a D_{a/b} + p_b D_A + p_a^2 p_b \end{aligned} \tag{11}$$

The trigenic disequilibria between one allele from marker **A** and two alleles from marker **B** is defined as

$$\begin{aligned} D_{aB} &= p_A^{BB} - p_B D_{ab} - p_B D_{a/b} - p_A D_B - p_A p_B^2 \\ &= -p_a^{BB} - p_B D_{ab} - p_B D_{a/b} + p_a D_B + p_a p_B^2 \\ &= -\frac{1}{2}p_A^{Bb} - \frac{1}{2}(p_B - p_b)D_{ab} \\ &\quad - \frac{1}{2}(p_B - p_b)D_{a/b} - p_a D_A + p_a p_B p_b \\ &= \frac{1}{2}p_a^{Bb} - \frac{1}{2}(p_B - p_b)D_{ab} \\ &\quad - \frac{1}{2}(p_B - p_b)D_{a/b} + p_a D_A - p_a p_B p_b \\ &= p_A^{bb} + p_b D_{ab} + p_b D_{a/b} - p_A D_A - p_A p_b^2 \\ &= -p_a^{bb} + p_b D_{ab} + p_b D_{a/b} + p_a D_A + p_a p_b^2 \end{aligned} \tag{12}$$

With diplotype frequencies, allele frequencies, HWD, gametic and non-gamete disequilibria, and trigenic disequilibria, we can derived the quadrigenic disequilibrium () between two alleles from marker **A** and two alleles from marker **B** (Weir, 1996). Analogous to Liu et al. (2006), we use a table (**Table 2**) to express the formulas for D_{AB}, from which it is clear that a full set of disequilibria can only be estimated from diplotype frequencies. In this article, D_{AB} will be estimated by using information from offspring genotypes. For clarity, we use small and capital letters to denote gamete and zygotic disequilibria, respectively. Conversely from **Table 2**, we can see that each of the diplotype frequencies can be expressed in terms of the allele frequencies (p_A, p_a and p_B, p_b), HWD coefficients

Table 2 | Expressions of quadrigenic disequilibrium D_{AB} in terms of genotypic configuration frequencies, allele frequencies and lower-order disequilibrium coefficients.

Frequency	1	$D_AD_B+D_{ab}^2+D_{a/b}^2$	D_A	D_B	D_{ab}	$D_{a/b}$	D_{Ab}	D_{aB}
P_{AA}^{BB}	$-p_A^2p_B^2$	-1	$-p_B^2$	$-p_A^2$	$-2p_Ap_B$	$-2p_Ap_B$	$-2p_B$	$-2p_A$
$-\frac{1}{2}P_{AA}^{Bb}$	$p_A^2p_Bp_b$	-1	p_Bp_b	$-p_A^2$	$-p_Ap_B+p_Ap_b$	$-p_Ap_B+p_Ap_b$	$-p_B+p_b$	$-2p_A$
P_{AA}^{bb}	$-p_A^2p_b^2$	-1	$-p_b^2$	$-p_A^2$	$2p_Ap_b$	$-2p_Ap_b$	$-2p_b$	$-2p_A$
$-\frac{1}{2}P_{Aa}^{BB}$	$p_Ap_ap_B^2$	-1	$-p_B^2$	p_Ap_a	$-p_Ap_B+p_ap_B$	$-\mathrm{p_ap_b}+p_ap_B$	$-2p_B$	$-p_A+p_a$
$\frac{1}{2}P_{Aa}^{Bb}$	$-p_Ap_ap_Bp_b$	-1	p_Bp_b	p_Ap_a	$-p_Ap_B-p_ap_b$	$p_Ap_b+p_ap_B$	$-p_B+p_b$	$-p_A+p_a$
$\frac{1}{2}P_{Aa}^{bB}$	$-p_Ap_ap_Bp_b$	-1	p_Bp_b	p_Ap_a	$p_Ap_b+p_ap_B$	$-\mathrm{p_ap_b}-p_ap_b$	$-p_B+p_b$	$-p_A+p_a$
$-\frac{1}{2}P_{Aa}^{bb}$	$p_Ap_Bp_b^2$	-1	$-p_b^2$	p_Ap_a	$p_Ap_b-p_ap_b$	$p_Ap_b-p_ap_b$	$2p_b$	$-p_A+p_a$
P_{aa}^{BB}	$-p_a^2p_B^2$	-1	$-p_B^2$	$-p_a^2$	$2p_ap_B$	$2p_ap_B$	$-2p_B$	$2p_a$
$-\frac{1}{2}P_{aa}^{Bb}$	$p_a^2p_Bp_b$	-1	p_Bp_b	$-p_a^2$	$p_ap_B-p_ap_b$	$p_ap_B-p_ap_b$	$-p_B+p_b$	$2p_a$
p_{aa}^{bb}	$-p_a^2p_b^2$	-1	p_b^2	$-p_a^2$	$-2p_ap_b$	$-2p_ap_b$	$2p_b$	$2p_a$

(D_A and D_B) and gametic (D_{ab}) and non-gametic disequilibria of different orders ($D_{a/b}$, D_{Ab}, D_{aB}, and D_{AB}).

MODEL FOR ESTIMATION

Based on the above description, it can be seen that the estimation of all disequilibrium parameters purely relies on the separation of two diplotypes underlying the double heterozygote, since all the other diplotypes are one to one corresponding to their genotypes and can be directly estimated from the data. In this section, we show that these two diplotypes can be separated using an open-pollinated (OP) design. For a monoecious plant, each plant may be OP randomly by both its own pollen and that from the natural pool. In a dioecious plant, a female plant is only pollinated by the pollen pool. To simplify our explanation, we focus on dioecious plants. A model for monoecious plants will be described elsewhere.

Let N_i denote the number of maternal plants which bear on genotype i (i=1,. . .,9) for the two SNPs considered and $N_i^{(j)}$ denote the number of offspring which have genotype j (j = 1,. . .,9) derived from maternal genotype i. **Table 3** gives the structure of genotypic data collected from $N. = \sum_{i=1}^{9} N_i$ random maternal plants and $N.^{(\cdot)} = \sum_{i=1}^{9}\sum_{j=1}^{9} N_i^{(j)}$ random offspring plants, respectively. For a given maternal genotype, a certain group of offspring genotypes is produced. **Table 4** gives genotype frequencies of offspring produced by any maternal genotype. For the OP design of dioecious plants, offspring genotypes are determined jointly by maternal gametes of the corresponding maternal plant and paternal gametes from the pollen pool. In the natural population, the frequencies of paternal gametes are expressed as p_{AB}, p_{Ab}, p_{aB}, and p_{ab}, respectively, which can be estimated from genotypic configuration frequencies by Eq. 4.

The frequencies of maternal gametes are dependent on the genotype of a maternal plant. If the maternal plant is a double heterozygote $AbBb$, then any genotype generated in the offspring will include a mixture of genotypes derived from gametes of the two underlying diplotypes of the double heterozygote. The proportions of mixture components are determined by two parameters, the recombination fraction (r) between the two markers, and the relative proportions (ϕ) of the two underlying diplotypes to the double heterozygote. For a maternal plant with genotype $AbBb$, there are two possible diplotypes, $AB|ab$ or $Ab|aB$, with relative frequencies

$$\phi = \frac{P_{Aa}^{Bb}}{P_{AaBb}}, 1-\phi = \frac{P_{Aa}^{bB}}{P_{AaBb}} \tag{13}$$

where P_{AaBb} is the frequency of the double heterozygote. Based on the definition of ϕ, we rewrite Eq. 4 to calculate the paternal gamete frequencies by

$$\begin{aligned} p_{AB} &= P_{AA}^{BB} + \frac{1}{2}\left(P_{AA}^{Bb} + P_{Aa}^{BB} + \phi P_{AaBb}\right) \\ p_{Ab} &= P_{AA}^{bb} + \frac{1}{2}\left(P_{AA}^{Bb} + P_{Aa}^{bb} + (1-\phi) P_{AaBb}\right) \\ p_{aB} &= P_{aa}^{BB} + \frac{1}{2}\left(P_{Aa}^{BB} + P_{aa}^{Bb} + (1-\phi) P_{AaBb}\right) \\ p_{ab} &= P_{aa}^{bb} + \frac{1}{2}\left(P_{Aa}^{bb} + P_{aa}^{Bb} + \phi P_{AaBb}\right). \end{aligned} \tag{14}$$

Maternal diplotypes, $AB|ab$ or $Ab|aB$, will produce four different maternal gametes, with relative frequencies depending on the recombination fraction (r) between the two markers:

$$\begin{array}{ll|llll} \text{Diplotype} & \text{Frequency} & AB & Ab & aB & ab \\ \hline AB|ab & \phi & \frac{1}{2}(1-r) & \frac{1}{2}r & \frac{1}{2}r & \frac{1}{2}(1-r) \\ Ab|aB & 1-\phi & \frac{1}{2}r & \frac{1}{2}(1-r) & \frac{1}{2}(1-r) & \frac{1}{2}r \\ \text{Overall} & 1 & \frac{1}{2}\theta & \frac{1}{2}(1-\theta) & \frac{1}{2}(1-\theta) & \frac{1}{2}\theta \end{array} \tag{15}$$

where we define $\theta = \phi(1-r) + r(1-\phi)$ and $1-\theta = \phi r + (1-r)(1-\phi)$. Thus, overall haplotype frequencies produced by the double heterozygote maternal are calculated as $\frac{1}{2}\theta$ for AB or ab and $\frac{1}{2}\theta$ for Ab and aB.

As seen from **Table 4**, the conditional genotype frequencies of offspring given maternal genotypes depend on parameter θ and paternal gamete frequencies. Since paternal gamete frequencies (14) are determined by unknown parameter ϕ and genotype/diplotype frequencies estimated from maternal genotypes (which can be thought of constants), these conditional probabilities are actually dependent on only θ and ϕ.

Table 3 | Data structure of two markers in the OP design.

Maternal family			Offspring genotype								
No.	**Genotype**	**Size**	**1** ***AABB*** ***AB|AB***	**2** ***AABb*** ***AB|Ab***	**3** ***AAbb*** ***Ab|Ab***	**4** ***AaBB*** ***AB|aB***	**5** ***AaBb*** ***AB|ab or Ab|aB***	**6** ***Aabb*** ***Ab|ab***	**7** ***aaBB*** ***aB|aB***	**8** ***aaBb*** ***aB|ab***	**9** ***aabb*** ***ab|ab***
1	*AABB*	N_1	$N_1^{(1)}$	$N_1^{(2)}$		$N_1^{(4)}$	$N_1^{(5)}$				
2	*AABb*	N_2	$N_2^{(1)}$	$N_2^{(2)}$	$N_2^{(3)}$	$N_2^{(4)}$	$N_2^{(5)}$	$N_2^{(6)}$			
3	*AAbb*	N_3		$N_3^{(2)}$	$N_3^{(3)}$		$N_3^{(5)}$	$N_3^{(6)}$			
4	*AaBB*	N_4	$N_4^{(1)}$	$N_4^{(2)}$		$N_4^{(4)}$	$N_4^{(5)}$		$N_4^{(7)}$	$N_4^{(8)}$	
5	*AaBb*	N_5	$N_5^{(1)}$	$N_5^{(2)}$	$N_5^{(3)}$	$N_5^{(4)}$	$N_5^{(5)}$	$N_5^{(6)}$	$N_5^{(7)}$	$N_5^{(8)}$	$N_5^{(9)}$
6	*Aabb*	N_6		$N_6^{(2)}$	$N_6^{(3)}$		$N_6^{(5)}$	$N_6^{(6)}$		$N_6^{(8)}$	$N_6^{(9)}$
7	*aaBB*	N_7				$N_7^{(4)}$	$N_7^{(5)}$		$N_7^{(7)}$	$N_7^{(8)}$	
8	*aaBb*	N_8				$N_8^{(4)}$	$N_8^{(5)}$	$N_8^{(6)}$	$N_8^{(7)}$	$N_8^{(8)}$	$N_8^{(9)}$
9	*aabb*	N_9					$N_9^{(5)}$	$N_9^{(6)}$		$N_9^{(8)}$	$N_9^{(9)}$

PARAMETER ESTIMATION WITH THE MCMC ALGORITHM

As have been clear above, only two parameters φ and θ determine the genotype frequencies of offspring (**Table 4**). The log-likelihood of a complete set of genotype data includes the two parts based on the maternal and offspring genotypes, respectively, expressed as

$$\log L\left(\phi,\theta\right) = \log L_M\left(\phi\right) + \log L_0\left(\phi,\theta\right). \tag{16}$$

We have

$$\log L_M\left(\phi\right) \propto N_1 \log P_{AABB} + \ldots + N_5 \log P_{AaBb} + \ldots + N_9 \log P_{aabb} \tag{17}$$

for the upper level of the log-likelihood that specifies the genotype distribution of maternal plants in the natural population, and

$$\begin{aligned}
&\log L_0(\phi,\theta) \propto \\
&+ N_1^{(1)} \log p_{AB} + N_1^{(2)} \log p_{Ab} + N_1^{(4)} \log p_{aB} + N_1^{(5)} \log p_{ab} \\
&+ \cdots \\
&+ N_5^{(1)} \log\left(\frac{1}{2}\theta p_{AB}\right) + N_5^{(2)} \log\left[\frac{1}{2}(\theta p_{Ab} + (1-\theta)p_{AB})\right] \\
&\qquad + N_5^{(3)} \log\left[\frac{1}{2}(1-\theta)p_{Ab}\right] \\
&+ N_5^{(4)} \log\left[\frac{1}{2}(\theta p_{aB} + (1-\theta)p_{AB})\right] + N_5^{(5)} \log\left[\frac{1}{2}\theta(P_{AB} + p_{ab})\right. \\
&\qquad \left. + \frac{1}{2}(1-\theta)(P_{Ab} + P_{aB})\right] \\
&+ N_5^{(6)} \log\left[\frac{1}{2}(\theta p_{Ab} + (1-\theta)P_{ab})\right] + N_5^{(7)} \log\left[\frac{1}{2}(1-\theta)p_{aB}\right] \\
&\qquad + N_5^{(8)} \log\left[\frac{1}{2}(1-\theta)p_{aB} + \theta P_{ab})\right] \\
&+ N_5^{(9)} \log\left(\frac{1}{2}\theta p_{ab}\right) \\
&+ \cdots \\
&+ N_9^{(5)} \log p_{AB} + N_9^{(6)} \log p_{Ab} + N_9^{(8)} \log p_{aB} + N_9^{(9)} \log p_{ab}
\end{aligned} \tag{18}$$

for the lower level of the log-likelihood that specifies the transmission of alleles from parents to offspring.

Since all the genotype frequencies of maternal plants can be directly estimated by the observed data, we can view the upper level of the log-likelihood as constant, thus only focus on the lower level $\log L_0(\phi,\theta)$, i.e., $\log L(\phi,\theta) \propto \log L_0(\phi\theta)$. Furthermore, $\log L_0(\phi,\theta)$ can be simplified as

$$\begin{aligned}
&\log L_0(\phi,\theta) \\
&\propto m_1 \log p_{AB} + m_2 \log p_{Ab} + m_3 \log p_{aB} + m_4 \log p_{ab} \\
&\quad + m_5 \log\theta + m_6 \log(1-\theta) + \\
&\quad + m_7 \log[\theta p_{ab} + (1-\theta)p_{AB}] + m_8 \log[\theta p_{aB} + (1-\theta)p_{AB}] \\
&\quad + m_9 \log[\theta p_{Ab} + (1-\theta)p_{ab}] + m_{10} \log[\theta p_{aB} + (1-\theta)p_{ab}] \\
&\quad + m_{11} \log[\theta(p_{AB} + p_{ab}) + (1-\theta)(p_{Ab} + p_{aB})].
\end{aligned}$$

where

$$\begin{aligned}
m_1 &= N_1^{(1)} + N_2^{(1)} + N_3^{(2)} + N_4^{(1)} + N_5^{(1)} + N_6^{(2)} + N_7^{(4)} \\
&\qquad + N_8^{(4)} + N_9^{(5)}, \\
m_2 &= N_1^{(2)} + N_2^{(3)} + N_3^{(3)} + N_4^{(2)} + N_5^{(3)} + N_6^{(3)} + N_7^{(5)} \\
&\qquad + N_8^{(6)} + N_9^{(6)}, \\
m_3 &= N_1^{(4)} + N_2^{(4)} + N_3^{(5)} + N_4^{(7)} + N_5^{(7)} + N_6^{(8)} + N_7^{(7)} \\
&\qquad + N_8^{(7)} + N_9^{(8)}, \\
m_4 &= N_1^{(5)} + N_2^{(6)} + N_3^{(6)} + N_4^{(8)} + N_5^{(9)} + N_6^{(9)} + N_7^{(8)} \\
&\qquad + N_8^{(9)} + N_9^{(9)}, \\
m_5 &= N_5^{(1)} + N_5^{(9)}, \qquad m_6 = N_5^{(3)} + N_5^{(7)}, \\
m_7 &= N_5^{(2)}, \qquad m_8 = N_5^{(4)}, \qquad m_9 = N_5^{(6)}, \\
m_{10} &= N_5^{(8)}, \qquad m_{11} = N_5^{(5)}.
\end{aligned} \tag{19}$$

Therefore, the posterior distribution of ϕ and θ based on all the other information can be obtained, respectively, by

Table 4 | Offspring genotype frequencies given each maternal genotype in the OP design.

Maternal family		Offspring genotype								
No.	**Genotype**	**1** ***AABB*** ***AB/AB***	**2** ***AABb*** ***AB/Ab***	**3** ***AAbb*** ***Ab/Ab***	**4** ***AaBB*** ***AB/aB***	**5** ***AaBb*** ***AB/ab Ab/aB***	**6** ***Aabb*** ***Ab/ab***	**7** ***aaBB*** ***aB/aB***	**8** ***aaBb*** ***aB/ab***	**9** ***aabb*** ***ab/ab***
1	*AABB*	p_{ab}	p_{Ab}		p_{aB}	p_{ab}				
2	*AABb*	$1/2p_{ab}$	$1/2(p_{AB}+p_{Ab})$	$1/2p_{Ab}$	$1/2p_{aB}$	$1/2(p_{ab}+p_{aB})$	$1/2p_{ab}$			
3	*AAbb*		p_{ab}	p_{Ab}		p_{aB}	p_{ab}			
4	*AaBB*	$1/2p_{AB}$	$1/2p_{Ab}$		$1/2(p_{AB}+p_{aB})$	$1/2(p_{ab}+p_{Ab})$		$1/2p_{aB}$	$1/2p_{ab}$	
5	*AaBb*	$1/2\theta p_{AB}$	$\frac{1}{2}(\theta p_{Ab}+\bar{\theta}p_{AB})$	$\frac{1}{2}\bar{\theta}p_{Ab}$	$\frac{1}{2}(\theta p_{aB}+\bar{\theta}p_{AB})$	$\frac{1}{2}(\theta p_{ab}+p_{AB})$ $+\frac{1}{2}\bar{\theta}(p_{aB}+p_{Ab})$	$\frac{1}{2}(\theta p_{Ab}+\bar{\theta}p_{ab})$	$\frac{1}{2}\bar{\theta}p_{aB}$	$\frac{1}{2}(\theta p_{aB}+\bar{\theta}p_{ab})$	$1/2\theta p_{ab}$
6	*Aabb*		$1/2p_{AB}$	$1/2p_{Ab}$		$1/2(p_{AB}+p_{aB})$	$1/2(p_{Ab}+p_{ab})$		$1/2p_{aB}$	$1/2p_{ab}$
7	*aaBB*				p_{ab}	p_{Ab}		p_{aB}	p_{ab}	
8	*aaBb*				$1/2p_{AB}$	$1/2(p_{AB}+p_{Ab})$	$1/2p_{Ab}$	$1/2p_{aB}$	$1/2(p_{aB}+p_{ab})$	$1/2p_{ab}$
9	*aabb*					p_{ab}	p_{Ab}		p_{aB}	p_{ab}

$\theta=\phi(1-r)+(1-\phi)r$, $\bar{\theta}=\phi r+(1-\phi)(1-r)$, and $\phi=P_{Aa}^{Bb}/P_{AaBb}$.

$$\begin{aligned}
&\log \pi(\phi|\theta, N) \\
&\propto m_1 \log p_{AB} + m_2 \log p_{Ab} + m_3 \log p_{aB} + m_4 \log p_{ab} \\
&\quad + m_7 \log\left[\theta p_{Ab} + (1-\theta)\, p_{AB}\right] + m_8 \log\left[\theta p_{aB} + (1-\theta)\, p_{AB}\right] \\
&\quad + m_9 \log\left[\theta p_{Ab} + (1-\theta)\, p_{ab}\right] + m_{10} \log\left[\theta p_{aB} + (1-\theta)\, p_{ab}\right] \\
&\quad + m_{11} \log\left[\theta\left(p_{AB} + p_{ab} + (1-\theta)(p_{Ab} + p_{aB})\right)\right]. \qquad (20)
\end{aligned}$$

$$\begin{aligned}
&\log \pi(\theta|\phi, N) \\
&\propto m_5 \log\theta + m_6 \log(1-\theta) + m_7 \log\left[\theta p_{Ab} + (1-\theta)\, p_{AB}\right] \\
&\quad + m_8 \log\left[\theta p_{aB} + (1-\theta)\, p_{AB}\right] + m_9 \log\left[\theta p_{Ab} + (1-\theta)\, p_{ab}\right] \\
&\quad + m_{10} \log\left[\theta p_{aB} + (1-\theta)\, p_{ab}\right] + m_{11} \log\left[\theta\left(p_{AB} + p_{ab}\right)\right. \\
&\quad \left. + (1-\theta)\left(p_{Ab} - p_{aB}\right)\right]. \qquad (21)
\end{aligned}$$

Notice that p_{AB}, p_{Ab}, p_{aB}, and p_{ab} are a function of ϕ according to equation (14). Then, within the Bayesian framework, the MCMC technique can be used to draw samples from the posterior distributions of ϕ and θ. We will use the Variable-at-a-Time Metropolis-Hastings algorithm described as follows:

1) Start with initial value $(\phi^{(0)},\theta^{(0)})$;
2) After getting $(\phi^{(n)},\theta^{(n)})$, update ϕ from $\phi^{(n)}$ to $\phi^{(n+1)}$ according to $\log\pi(\phi|\theta^{(n)},N)$ using proposal Unif(0,1): Generate ϕ^* from Unif(0,1) and accept it, i.e., set $\phi^{(n+1)} = \phi^*$ with probability α, where $\log\alpha = \min\{0, \log\pi(\phi^*|\theta^{(n)},N) - \log\pi(\phi^{(n)}|\theta^{(n)},N)\}$; otherwise, set $\phi^{(n+1)} = \phi^{(n)}$;
3) After getting $(\phi^{(n+1)} = \theta^{(n)})$, update θ from $\theta^{(n)}$ to $\theta^{(n+1)}$ according to $\log\pi(\theta|\phi^{(n+1)},N)$ using proposal Unif(0,1): Generate θ^* from Unif(0,1) and accept it, i.e., set $\theta^{(n+1)} = \theta^*$ with probability β, where $\log\beta = \min\{0, \log\pi(\theta * |\phi^{(\mathrm{n}+1)},N) - \log\pi(\theta^{(n)}|\phi^{(n+1)},N)\}$; otherwise, set $\theta^{(n+1)} = \theta^{(n)}$;
4) Repeat (2) and (3) n times, where n is the number of iterations.
5) After burning in the first few sampled ϕ and θ, get the mean values of the remaining samples as the respective estimates. The estimate of r can be obtained by

$$r = \frac{\theta - \phi}{1 - 2\phi} \text{ from } \theta = \phi\,(1-r) + r\,(1-\phi)\,.$$

Since the values of r, ϕ and, hence, θ have restricted domains as follows:

$$0 \le r \le 0.5; 0 \le \phi \le 1; 0 \le \theta \le 1,$$

we choose Uniform (0,1) as a proposal distribution when updating both ϕ and θ.

With the estimate of ϕ, we can separate two diplotypes of the double heterozygote. Therefore, all the frequencies and disequilibria are obtained from Eqs 1 to 12 and **Table 2**. Also, we provide an estimate of the recombination fraction.

COMPUTER SIMULATION

The statistical properties of the new model were investigated through simulation studies. The values of zygotic disequilibria used to simulate marker data for the OP design were chosen from their spaces which were shown in Liu et al. (2006). Because of the limitation of a population-based sampling design, Liu et al. (2006) was not able to estimate all types of zygotic disequilibria. By collapsing gametic and non-gametic linkage disequilibria, leading to a composite quadrigenic disequilibrium rather than the quadrigenic disequilibrium as defined in **Table 2**, they provided a reduced model for disequilibrium analyses. In a comparison with the traditional gametic linkage disequilibrium model, Liu et al.'s reduced model was found to be more general and cover the results from the former. It is expected that our model covers Liu et al.'s model because we do not need to combine gametic and non-gametic linkage disequilibria.

The simulation uses three different scenarios. The first assumes different relative proportions of two diplotypes for the double heterozygote (ϕ) by fixing the recombination fraction. In order to obtain different ϕ values, we need to adjust the zygotic disequilibria. Specifically, we have three combinations:

(1) $\phi = 0.87$ and $D_{ab} = 0.1$, $D_{a/b} = D_{Ab} = D_{aB} = D_{AB} = 0.01$
(2) $\phi = 0.12$ and $D_{a/b} = 0.1$, $D_{ab} = D_{Ab} = D_{aB} = D_{AB} = 0.01$
(3) $\phi = 0.5$ and $D_{ab} = D_{a/b} = 0.02$, $D_{Ab} = D_{aB} = 0.03$, $D_{AB} = 0.01$,

in which $r = 0.2$, $p_A = 0.5$, $p_B = 0.6$, $D_A = D_B = 0.05$, and the true value of θ can be calculated given ϕ and r by equation (15). The second is to change the recombination fraction, i.e., (1) $r = 0.3$, (2) $r = 0.05$, and (3) $r = 0.5$, with the other parameters fixed, i.e., $\phi = 0.87$, $p_A = 0.5$, $p_A = 0.6$, $D_A = D_B = 0.05$, $D_{ab} = 0.01$, $D_{a/b} = D_{Ab} = D_{aB} = D_{AB} = 0.01$. The third is about sampling strategies, (1) 1000 maternal plants × 9 seeds per family and (2) 200 maternal plants × 49 seeds per family, producing the same total size of samples, in which $\phi = 0.87$, $r = 0.2$, $p_A = 0.5$, $p_B = 0.6$, $D_A = D_B = 0.05$, $D_{ab} = 0.1$, $D_{a/b} = D_{Ab} = D_{aB} = D_{AB} = 0.01$. For the first two scenarios, we use 1000 × 9 strategy.

For the simulated data, we estimated all genotype frequencies from maternal genotype data, which were used to estimate p_A, p_B, D_A, D_B, D_{Ab}, and D_{aB} with equations (1), (2), (3), (6), (7), (8), (11), and (12), respectively. However, D_{ab}, $D_{a/b}$, and D_{AB} can be estimated only after ϕ is estimated. **Tables 5–7** give the results from simulated data under three different scenarios. In each case, 1000 simulation replicates were performed to get the means and standard deviations of the estimates. Each estimate was based on 1000 iterations after burns-in during the MCMC procedure. In general, all parameters can be estimated reasonably from our model.

In scenario 1, uneven allocations of two diplotypes for the double heterozygote help the estimates of the recombination fraction (**Table 5**). If these two diplotypes are equally allocated, i.e., $\phi = 0.5$, we could not give a good estimate of the recombination fraction because in this case θ is not dependent on r, making r unidentifiable. As shown in Li and Wu (2009), a three-locus analysis can overcome this problem. In scenario 2, we found that ϕ and all set of disequilibria can be precisely estimated, no matter whether $r = 0.5$ or not (**Table 6**). This is because the MCMC procedure treats ϕ and θ as two separate parameters. The results of scenario 3 help to determine an optimal sampling strategy. As expected, sampling more maternal plants increases the precision of parameter estimation (**Table 7**).

Table 5 | Estimates of parameters and their standard deviations (in parentheses) based on the MCMC algorithm for the data simulated under scenario 1.

	ϕ	θ	r	p_A	p_B	D_A	D_B	Dab	$D_{a/b}$	D_{Ab}	D_{aB}	D_{AB}
DESIGN 1												
True	0.87	0.72	0.2	0.5	0.6	0.05	0.05	0.1	0.01	0.01	0.01	0.01
Estimate	0.871	0.724	0.192	0.500	0.599	0.050	0.099	0.011	0.010	0.010	0.010	0.010
SD	(0.036)	(0.021)	(0.049)	(0.009)	(0.008)	(0.010)	(0.007)	(0.003)	(0.009)	(0.003)	(0.003)	(0.002)
DESIGN 2												
True	0.12	0.28	0.2	0.5	0.6	0.05	0.05	0.01	0.1	0.01	0.01	0.01
Estimate	0.122	0.276	0.196	0.500	0.600	0.050	0.050	0.010	0.100	0.010	0.010	0.010
SD	(0.048)	(0.021)	(0.059)	(0.009)	(0.011)	(0.007)	(0.008)	(0.003)	(0.007)	(0.003)	(0.003)	(0.003)
DESIGN 3												
True	0.5	0.5	0.2	0.5	0.6	0.05	0.05	0.02	0.02	0.03	0.03	0.01
Estimate	0.496	0.501	0.374	0.500	0.600	0.050	0.050	0.020	0.021	0.030	0.030	0.010
SD	(0.053)	(0.022)	(7.368)	(0.012)	(0.012)	(0.008)	(0.008)	(0.003)	(0.010)	(0.003)	(0.003)	(0.002)

Table 6 | Estimates of parameters and their standard deviations (in parentheses) based on the MCMC algorithm for the data simulated under scenario 2.

	ϕ	θ	r	p_A	p_B	D_A	D_B	D_{ab}	$D_{a/b}$	D_{Ab}	D_{aB}	D_{AB}
DESIGN 1												
True	0.87	0.65	0.3	0.5	0.6	0.05	0.05	0.1	0.01	0.01	0.01	0.01
Estimate	0.871	0.640	0.295	0.500	0.599	0.050	0.050	0.100	0.011	0.010	0.010	0.010
SD	(0.045)	(0.018)	(0.040)	(0.012)	(0.012)	(0.008)	(0.008)	(0.003)	(0.010)	(0.003)	(0.003)	(0.002)
DESIGN 2												
True	0.87	0.84	0.05	0.5	0.6	0.05	0.05	0.1	0.01	0.01	0.01	0.01
Estimate	0.875	0.837	0.044	0.500	0.600	0.050	0.050	0.100	0.010	0.010	0.010	0.010
SD	(0.045)	(0.016)	(0.067)	(0.012)	(0.012)	(0.008)	(0.007)	(0.003)	(0.010)	(0.003)	(0.003)	(0.002)
DESIGN 3												
True	0.87	0.5	0.5	0.5	0.6	0.05	0.05	0.1	0.01	0.01	0.01	0.01
Estimate	0.874	0.499	0.501	0.500	0.600	0.049	0.050	0.100	0.010	0.010	0.010	0.010
SD	(0.046)	(0.018)	(0.025)	(0.012)	(0.012)	(0.008)	(0.007)	(0.003)	(0.010)	(0.003)	(0.003)	(0.002)

Table 7 | Estimates of parameters and their standard deviations (in parentheses) based on the MCMC algorithm for the data simulated under scenario 3.

	ϕ	θ	r	p_A	p_B	D_A	D_B	D_{ab}	$D_{a/b}$	D_{Ab}	D_{aB}	D_{AB}
SAMPLING STRATEGY 1: 1000 × 9												
True	0.87	0.72	0.2	0.5	0.6	0.05	0.05	0.1	0.01	0.01	0.01	0.01
Estimate	0.871	0.724	0.192	0.500	0.599	0.050	0.099	0.011	0.010	0.010	0.010	0.010
SD	(0.036)	(0.011)	(0.049)	(0.009)	(0.008)	(0.010)	(0.007)	(0.003)	(0.009)	(0.003)	(0.003)	(0.002)
SAMPLING STRATEGY 2: 200 × 49												
True	0.87	0.72	0.2	0.5	0.6	0.05	0.05	0.1	0.01	0.01	0.01	0.01
Estimate	0.866	0.726	0.154	0.500	0.599	0.049	0.050	0.098	0.011	0.010	0.010	0.010
SD	(0.085)	(0.018)	(0.500)	(0.028)	(0.027)	(0.027)	(0.017)	(0.004)	(0.021)	(0.006)	(0.006)	(0.004)

DISCUSSION

The structural study of linkage disequilibrium helps to understand the genetic variation and evolution of populations and facilitates the positional cloning of genes underlying common complex diseases (Lewontin, 1964, 1988; Hedrick, 1987; Weir, 1996; Kruglyak, 1999). The current approaches for estimating linkage disequilibria rely on the assumption that the population under consideration is randomly mating, following Hardy–Weinberg equilibrium (HWE). However, many populations may be founded by a small number of ancestors and/or are frequently under evolutionary pressure, such as mutation, genetic drift, population admixture, and structure (Lynch and Walsh, 1998). For those populations, HWE may be violated. We need a new analysis that relaxes the random mating assumption. Weir (1996) introduced the concept of zygotic association or zygotic disequilibrium that specify the disequilibria between different loci in a non-equilibrium population. Part of these disequilibria was used by Liu et al. (2006) to examine the extent and distribution of zygotic disequilibria across the canine genome.

In this article, we have for the first time developed a new statistical model for estimating a complete set of zygotic disequilibria, including (1) Hardy–Weinberg disequilibria at each locus, (2) gametic disequilibrium (including two alleles in the same gamete, each from a different locus), (3) non-gametic disequilibrium (including two alleles in different gametes, each from a different locus), (4) trigenic disequilibrium (including a zygote at one locus and an allele at the other), and (5) quadrigenic disequilibrium (including two zygotes each from a different locus). This model is based on an open-pollinated (OP) design by sampling multiple maternal plants and their offspring. The major advantage of this design lies in its power to separate two diplotypes of a double heterozygote (specified by a proportion φ), thus retrieving the lost information of a traditional population-based sampling strategy.

Under the OP design setting, a full log-likelihood of diplotype frequencies from the maternal and offspring generations was formulated in terms of two unique parameters, ϕ and recombination fraction *r*. An MCMC procedure was then implemented to estimate these two parameters from which a full set of zygotic disequilibria and the recombination fraction are estimated. Extensive simulation studies have been performed to test the statistical behavior of the new model by considering a range of disequilibrium and recombination fraction as well as different sampling strategies. In general, all the parameters can be estimated with reasonable accuracy and precision.

As a first attempt to relax the Hardy–Weinberg equilibrium assumption that has been widely accepted for population genetic studies of many decades, our model will reshape the fundamental theory of this field. With an increasing availability of genetic data due to the rapid development of genotyping technologies, this model will show its increasing implications and likelihood for uncovering new discoveries related to population genetics. The model can be extended to accommodate various situations in the following aspects. First, by integrating selfing rates into the model, we can develop a similar design for monoecious plants in which the simultaneous occurrence of female and male flowers allows selfing. Second, a multilocus model including more than two markers should be developed. This will not only increase the power of the model by estimating genetic interference, but also overcome the problem of estimating the recombination fraction when φ is equal to 0.5. Third, the model can be integrated with quantitative traits to map their underlying QTLs and genetic interactions (see Wu et al., 2002). All these extensions will make our model more useful in practice, ultimately resolving difficult challenges in population and quantitative genetic studies.

ACKNOWLEDGMENTS

This work is partially supported by NSF/IOS-0923975 to Rongling Wu and NIDA, NIH grants R21 DA024260 and R21 DA024266 to Runze Li. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or the NIH.

REFERENCES

Barton, N. H., and Gale, K. S. (1993). "Genetic analysis of hybrid zones," in *Hybrid Zones and the Evolutionary Process*, eds R. G. Harrison and J. Price (Oxford: Oxford University Press), 13–45.

Bennett, J. H., and Binet, F. E. (1956). Association between Mendelian factors with mixed selfing and random mating. *Heredity* 10, 51–55.

Charlesworth, B. (1991). The evolution of sex chromosomes. *Science* 251, 1030–1033.

Haldane, J. B. S. (1949). The association of characters as a result of inbreeding and linkage. *Ann. Eugen.* 15, 15–23.

Hedrick, P. W. (1987). Gametic disequilibrium measures: proceed with caution. *Genetics* 117, 331–341.

Hou, W., Liu, T., Li, Y., Li, Q., Li, J. H., Das, K., and Wu, R. L. (2009). Multilocus genomics of outcrossing populations. *Theor. Popul. Biol.* 76, 68–76.

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

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

Lewontin, R. C. (1988). On measures of gametic disequilibrium. *Genetics* 120, 849–852.

Li, J. H., Li, Q., Hou, W., Han, K., Li, Y., Wu, S., Li, Y. C., and Wu, R. L. (2009). An algorithmic model for constructing a linkage and linkage disequilibrium map in open-pollinated progeny populations. *Genet. Res.* 91, 9–21.

Li, Q., and Wu, R. L. (2009). A multilocus model for constructing a linkage disequilibrium map in human populations. *Stat. Appl. Genet. Mol. Biol.* 8, article 18.

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

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

Lou, X. Y., Casella, G., Littell, R. C., Yang, M. C. K., Johnson, J. A., and Wu, R. L. (2003). A haplotype-based algorithm for multilocus linkage disequilibrium mapping of quantitative trait loci with epistasis. *Genetics* 163, 1533–1548.

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

Slatkin, M. (2008). Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* 9, 477–485.

Weir, B. (1996). *Genetic Data Analysis II*. Sunderland, MA: Sinauer Associates.

Wu, R. L., Ma, C. X., and Casella, G. (2002). Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations. *Genetics* 160, 779–792.

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

Wu, S., Yang, J., and Wu, R. L. (2010). Mapping quantitative trait loci in a non-equilibrium population. *Stat. Appl. Genet. Mol. Biol.* 9, article 32.

Yang, R. C. (2000). Zygotic associations and multilocus statistics in a nonequilibrium diploid population. *Genetics* 155, 1449–1458.

Yang, R. C. (2002). Analysis of multilocus zygotic associations. *Genetics* 161, 435–445.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 September 2011; paper pending published: 31 October 2011; accepted: 24 April 2012; published online: 02 July 2012.

Citation: Liu J, Wang Z, Wang Y, Li R and Wu R (2012) Model and algorithm for linkage disequilibrium analysis in a non-equilibrium population. Front. Gene. ***3****:78. doi: 10.3389/fgene.2012.00078*

This article was submitted to Frontiers in Statistical Genetics and Methodology, a specialty of Frontiers in Genetics.