Genet. Res., Camb. (2009), **91**, *pp.* 9–21. © 2009 Cambridge University Press
doi:10.1017/S0016672308009932 Printed in the United Kingdom

An algorithmic model for constructing a linkage and linkage disequilibrium map in outcrossing plant populations

JIAHAN LI[1]*, QIN LI[1]*, WEI HOU[2]*, KUN HAN[3]*, YAO LI[1], SONG WU[1], YANCHUN LI[3] AND RONGLING WU[1,3,4]*†

[1] *Department of Statistics, University of Florida, Gainesville, FL 32611, USA*
[2] *Department of Epidemiology and Health Policy Research, University of Florida, Gainesville, FL 32611, USA*
[3] *School of Forestry and Biotechnology, Zhejiang Forestry University, Lin'an, Zhejiang 311300, People's Republic of China, and*
[4] *Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA*

(*Received 8 September 2008 and in revised form 28 October 2008*)

Summary

A linkage–linkage disequilibrium map that describes the pattern and extent of linkage dis-equilibrium (LD) decay with genomic distance has now emerged as a viable tool to unravel the genetic structure of population differentiation and fine-map genes for complex traits. The prerequisite for constructing such a map is the simultaneous estimation of the linkage and LD between different loci. Here, we develop a computational algorithm for simultaneously estimating the recombination fraction and LD in a natural outcrossing population with multilocus marker data, which are often estimated separately in most molecular genetic studies. The algorithm is founded on a commonly used progeny test with open-pollinated offspring sampled from a natural population. The information about LD is reflected in the co-segregation of alleles at different loci among parents in the population. Open mating of parents will reveal the genetic linkage of alleles during meiosis. The algorithm was constructed within the polynomial-based mixture framework and implemented with the Expectation–Maximization (EM) algorithm. The by-product of the derivation of this algorithm is the estimation of outcrossing rate, a parameter useful to explore the genetic diversity of the population. We performed computer simulation to investigate the influences of different sampling strategies and different values of parameters on parameter estimation. By providing a number of testable hypotheses about population genetic parameters, this algorithmic model will open a broad gateway to understand the genetic structure and dynamics of an outcrossing population under natural selection.

1. Introduction

Linkage disequilibrium (LD) refers to the non-random association of alleles at different loci in the genome. Historically, LD analysis was developed to study the genetic structure and diversity of natural populations (Lewontin, 1964; Hill, 1974; Hedrick, 1987; Weir, 1996). In recent years, there has been a dramatic increase of interest in utilizing LD to infer the evolutionary history and process of human populations (Reich *et al*., 2001; Ardlie *et al*., 2002; Dawson *et al*., 2002; Gabriel *et al*., 2002) and to identify genes for disease or yield traits by association analyses with DNA-based markers (Remington *et al*., 2001; Ardlie *et al*., 2002). The efficacy of LD analysis in population genetic studies and gene mapping depends on the level of LD in the population studied, its distribution and heterogeneity across the genome, and its relationship with genetic or geographic distances (Farnir *et al*., 2000; McRae *et al*., 2002; Liu *et al*., 2006). The pattern of LD decay is well known in human populations, where the relationship between the LD and physical distance is graphically elucidated to infer the origin and evolution of humans (Tishkoff *et al*., 1996, 2001; Tishkoff & Williams, 2002).

The pattern of LD decay with genetic distance (measured in terms of the recombination fraction) can be used to characterize the genetic structure and

† Corresponding author. Department of Statistics, University of Florida, Gainesville, FL 32611, USA. Tel: (352)392-3806. Fax. (352)392-8555. e-mail: rwu@stat.ufl.edu
* These authors contributed equally to this work.

dynamics of populations. This needs the simultaneous measures of the LD and recombination fraction between the same pair of loci. However, the estimation of these two parameters is usually based on different genetic designs; i.e. the estimation of the recombination fraction relies on a segregating pedigree, whereas the estimation of LD needs a random sample drawn from a natural population. More recently, several designs have been proposed to jointly measure the linkage and linkage disequilibrium for natural populations (Wu & Zeng, 2001) and domestic animals (Georges, 2007). Simultaneous estimation of LD and the recombination fraction can avoid false positive results (spurious LD) when LD is used to fine-map genes for complex traits given the frequent occurrence of LD between distantly spaced loci or unlinked loci.

Wu & Zeng (2001) proposed an open-pollinated (OP) design for population genetic studies of forest trees with molecular markers. For most forest tree species, seeds from a single mother tree are derived from the open pollination of unknown fathers from the pollen pool. By collecting OP seeds from a sample of individual trees in a natural population, Wu & Zeng's design did not take into account the hermaphroditical nature of a tree species in which both sexes exist on the same individual, and thus its seeds may be derived from both selfing and outcrossing pollination. Self-fertilization is thought to affect diversity by reduced effective population size and reduced genome-wide effective recombination rates, both due to increased homozygosity, elevated isolation among individuals and subpopulations induced by inbreeding (Charlesworth, 2003). Consequently, a predominantly selfing mode of reproduction may be expected to lead to low polymorphism, extensive LD and high population subdivision (Nordborg, 2000; Ingvarsson, 2002, 2005). These predictions can be tested by simultaneous estimation of the outcrossing rate, recombination fraction and LD. Also, the OP design proposed by Wu & Zeng (2001) did not incorporate a procedure for estimating the diplotype of heterozygous trees from which seeds are sampled. In this paper, we extend Wu & Zeng's OP progeny design to better understand the genetic structure of a natural population by simultaneously estimating multiple population genetic parameters with molecular markers. Simulation studies were performed to examine the statistical behaviour of the model.

2. Model

(i) *Sampling and genotyping strategy*

Suppose there is a natural population at Hardy–Weinberg equilibrium (HWE) for a dioecious plant species. Each plant in the population is OP by its own pollen (selfing) and randomly by the pollen from other individuals (outcrossing). Thus, seeds produced by each plant include a mix of offspring due to selfing and outcrossing pollination. We will randomly sample a set of maternal plants and further randomly collect a sample of seeds from each sampled plant. Because the fathers of seeds from a sampled maternal plant are unknown, this sampling strategy will generate a set of half-sib families. The collected seeds (embryos) are germinated into seedlings. DNA samples are taken from maternal plants and their offspring derived from the seeds for marker analysis.

A panel of molecular markers is typed to examine population genetic properties by estimating the recombination fraction, LD and outcrossing rate. Consider two markers, each with two alleles, 1 and 0, which are generally denoted by i for the first marker ($i=1, 0$) and j for the second marker ($j=1, 0$). Different alleles at each marker unite to form four gametes, whose frequencies in the population are expressed as

$$\begin{aligned} p_{11} &= pq + D && \text{for gamete 11,} \\ p_{10} &= p(1-q) - D && \text{for gamete 10,} \\ p_{01} &= (1-p)q - D && \text{for gamete 01,} \\ p_{00} &= (1-p)(1-q) + D && \text{for gamete 00,} \end{aligned} \tag{1}$$

which sum to one, where p and $1-p$ are the frequencies of two alleles, 1 and 0, for the first marker, q and $1-q$ are the frequencies of two alleles for the second marker, and D is the degree of gametic LD between the two markers. It is assumed that there is no sex-specific difference in gamete frequencies, allele frequencies and LD in the population.

Among the sampled maternal plants, there are nine genotypes for the two markers considered, generally expressed as $ii'jj'$ ($i \geqslant i'=1, 0$; $j \geqslant j'=1, 0$). Let $N_{ii'jj'}$ denote the number of maternal plants with genotype $ii'jj'$. Under the assumption of HWE, the frequency of a diplotype is the product of the frequencies of the gametes that form the diplotype. By collapsing those diplotypes that are observed as the same genotype, the frequencies of genotypes are generally expressed as

$$P_{ii'jj'} = \begin{cases} p_{ij}^2 & \text{for } i=i' \text{ and } j=j', \\ p_{ij}p_{ij'} + p_{ij'}p_{ij} & \text{for } i=i' \text{ and } j\neq j', \\ p_{ij}p_{i'j} + p_{i'j}p_{ij} & \text{for } i\neq i' \text{ and } j=j', \\ p_{ij}p_{i'j'} + p_{i'j'}p_{ij} & \\ \quad + p_{ij'}p_{i'j} + p_{i'j}p_{ij'} & \text{for } i\neq i' \text{ and } j\neq j'. \end{cases} \tag{2}$$

Table 1 gives the genotype frequencies of the maternal plants for the two markers in terms of haplotype or diplotype frequencies calculated with equation (2).

The seeds collected from each sampled maternal plant are typed for the two markers so that the genotype of each offspring can be known. The same offspring genotype from the same maternal genotype are mixed up. Let $N_{ii'jj'}^{ll'rr'}$ be the mixed number of offspring with genotype $ll'rr'$ ($l \geqslant l'=1, 0$; $r \geqslant r'=1, 0$) collected from $N_{ii'jj'}$ maternal plants with genotype $ii'jj'$. From

https://doi.org/10.1017/S0016672308009932 Published online by Cambridge University Press

Table 1. *Diplotype and genotype frequencies of two markers, A and B, in the offspring population through outcrossing and selfing pollination*

Maternal				Offspring										
				11/11	11/10	11/00	10/11	10/10			10/00	00/11	00/10	00/00
Genotype	Diplotype	Frequency	Proportion	11|11	11|10	10|10	11|01	11|00	+	10|01	10|00	01|01	01|00	00|00
11/11	11|11	p_{11}^2	$1-w$	1	0	0	0	0	+	0	0	0	0	0
			w	p_{11}	p_{10}	0	p_{01}	p_{00}	+	0	0	0	0	0
11/10	11|10	$p_{11}p_{10}$	$1-w$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	0	0	+	0	0	0	0	0
			w	$\frac{1}{2}p_{11}$	$\frac{1}{2}(p_{11}+p_{10})$	$\frac{1}{2}p_{10}$	$\frac{1}{2}p_{01}$	$\frac{1}{2}p_{00}$	+	$\frac{1}{2}p_{01}$	$\frac{1}{2}p_{00}$	0	0	0
11/00	10|10	p_{10}^2	$1-w$	0	p_{11}	p_{10}	0	0	+	p_{01}	p_{00}	0	0	0
			w	0	0	1	0	0	+	0	0	0	0	0
10/11	11|01	$2p_{11}p_{10}$	$1-w$	$\frac{1}{4}$	0	0	$\frac{1}{2}$	0	+	0	0	$\frac{1}{4}$	0	0
			w	$\frac{1}{2}p_{11}$	$\frac{1}{2}p_{10}$	0	$\frac{1}{2}(p_{11}+p_{01})$	$\frac{1}{2}p_{00}$	+	$\frac{1}{2}p_{10}$	0	$\frac{1}{2}p_{01}$	$\frac{1}{2}p_{00}$	0
10/10	11|00	$2p_{11}p_{00}$	$1-w$	$\frac{1}{4}\bar{r}^2$	$\frac{1}{2}r\bar{r}$	$\frac{1}{4}r^2$	$\frac{1}{2}r\bar{r}$	$\frac{1}{2}\bar{r}^2$	+	$\frac{1}{2}r^2$	$\frac{1}{2}r\bar{r}$	$\frac{1}{4}r^2$	$\frac{1}{2}r\bar{r}$	$\frac{1}{4}\bar{r}^2$
	10|01	$p_{10}p_{01}$	w	$\frac{1}{2}\bar{r}p_{11}$	$\frac{1}{2}(rp_{11}+\bar{r}p_{10})$	$\frac{1}{2}rp_{10}$	$\frac{1}{2}(rp_{11}+\bar{r}p_{01})$	$\frac{1}{2}\bar{r}(p_{11}+p_{00})$	+	$\frac{1}{2}r(p_{01}+p_{10})$	$\frac{1}{2}(\bar{r}p_{10}+rp_{00})$	$\frac{1}{2}rp_{01}$	$\frac{1}{2}(\bar{r}p_{01}+rp_{00})$	$\frac{1}{2}\bar{r}p_{00}$
			$1-w$	$\frac{1}{4}r^2$	$\frac{1}{2}r\bar{r}$	$\frac{1}{4}\bar{r}^2$	$\frac{1}{2}r\bar{r}$	$\frac{1}{2}r^2$	+	$\frac{1}{2}\bar{r}^2$	$\frac{1}{2}r\bar{r}$	$\frac{1}{4}\bar{r}^2$	$\frac{1}{2}r\bar{r}$	$\frac{1}{4}r^2$
			w	$\frac{1}{2}rp_{11}$	$\frac{1}{2}(\bar{r}p_{11}+rp_{10})$	$\frac{1}{2}\bar{r}p_{10}$	$\frac{1}{2}(\bar{r}p_{11}+rp_{01})$	$\frac{1}{2}r(p_{11}+p_{00})$	+	$\frac{1}{2}\bar{r}(p_{10}+p_{01})$	$\frac{1}{2}(rp_{10}+\bar{r}p_{00})$	$\frac{1}{2}\bar{r}p_{01}$	$\frac{1}{2}(rp_{01}+\bar{r}p_{00})$	$\frac{1}{2}rp_{00}$
10/00	10|00	$2p_{10}p_{00}$	$1-w$	0	0	$\frac{1}{4}$	0	0	+	0	$\frac{1}{2}$	0	0	$\frac{1}{4}$
			w	0	$\frac{1}{2}p_{11}$	$\frac{1}{2}p_{10}$	0	$\frac{1}{2}p_{11}$	+	$\frac{1}{2}p_{01}0$	$\frac{1}{2}(p_{10}+p_{00})$	0	$\frac{1}{2}p_{01}$	$\frac{1}{2}p_{00}$
00/11	01|01	p_{01}^2	$1-w$	0	0	0	0	0	+	0	0	1	0	0
			w	0	0	0	p_{11}	0	+	p_{10}	0	p_{01}	p_{00}	0
00/10	01|00	$2p_{01}p_{00}$	$1-w$	0	0	0	0	0	+	0	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
			w	0	0	0	$\frac{1}{2}p_{11}$	$\frac{1}{2}p_{11}$	+	$\frac{1}{2}p_{10}$	$\frac{1}{2}p_{10}$	$\frac{1}{2}p_{01}$	$\frac{1}{2}(p_{01}+p_{00})$	$\frac{1}{2}p_{00}$
00/00	00|00	p_{00}^2	$1-w$	0	0	0	0	p_{11}	+	0	p_{10}	0	p_{01}	p_{00}
			w	0	0	0	0	0	+	0	0	0	0	1

Table 2. *Two possible diplotypes of a maternal plant of double heterozygote and the frequencies of its four gametes for two markers*

Maternal diplotype	Relative proportion	Gamete			
		11	10	01	00
11|00	ϕ	$\frac{1}{2}(1-r)$	$\frac{1}{2}r$	$\frac{1}{2}r$	$\frac{1}{2}(1-r)$
10|01	$1-\phi$	$\frac{1}{2}r$	$\frac{1}{2}(1-r)$	$\frac{1}{2}(1-r)$	$\frac{1}{2}r$

observed offspring genotypes, we will estimate key population genetic parameters that define population structure and organization.

(ii) *Offspring structure*

Each sampled maternal plant undergoes meiosis to produce male and female gametes. For those double homozygotes at the two markers considered, only one gamete type is yielded. The plants which are heterozygous only for one marker produce two types of gametes with equal frequency. For the double heterozygote plants, there are four possible types of gametes: 11, 10, 01 and 00. This type of plant has two possible diplotypes 11|00 and 01|10, which will produce different gamete frequencies expressed as a function of the recombination fraction (r) (Table 2). Of all double heterozygotes, there is a relative proportion of

$$\phi = \frac{p_{11}p_{00}}{p_{11}p_{00}+p_{10}p_{01}}$$

for diplotype 11|00, and

$$\bar{\phi} = 1-\phi = \frac{p_{10}p_{01}}{p_{11}p_{00}+p_{10}p_{01}}$$

for diplotype 10|01.

Each female gamete produced by a maternal plant unites at random with its own male gamete to form a selfing offspring, or with a gamete from the pollen pool to form an outcrossing offspring. Table 1 lists the frequencies of two-marker diplotypes (and therefore genotypes) in the selfing and outcrossing offspring populations produced by possible maternal genotypes. The pollen pool that contributes to the outcrossing seeds contains four male gametes, 11, 10, 01 and 00, whose frequencies are defined by p_{11}, p_{10}, p_{01} and p_{00}, respectively. Let w be the outcrossing rate of the plant measured by the proportion of its offspring that are generated through fertilization by pollens of other plants in the population. Thus, the selfing rate of the plant that receives its own pollen to pollinate is $\bar{w}=1-w$.

Although marker genotypes of offspring sampled from a given maternal genotype can be observed, the mechanisms of genotype formation are unknown. The formation of progeny genotypes includes four mechanisms:

(1) **Pollination behaviour:** The same progeny genotype can be derived from the selfing or outcrossing of a maternal plant. Let $P_{ii'jj'}^{ll'rr'}$ denote the overall frequency of offspring genotype $ll'rr'$ derived from maternal genotype $ii'jj'$, which is generally expressed, by considering all possible mechanisms of genotype formation, as

$$P_{ii'jj'}^{ll'rr'} = \bar{w}S_{ii'jj'}^{ll'rr'} + wO_{ii'jj'}^{ll'rr'}, \tag{3}$$

where $S_{ii'jj'}^{ll'rr'}$ and $O_{ii'jj'}^{ll'rr'}$ are the frequencies of offspring genotype $ll'rr'$ derived from maternal genotype $ii'jj'$ due to the maternal plant's selfing and outcrossing pollination, respectively (Table 1).

(2) **Sex origin of a gamete:** The same progeny genotype may be due to reciprocal combinations between two gametes from male and female sides. For example, a maternal genotype 11/10 yields two female gametes, 11 and 10. When they unite with male gametes 10 and 11, respectively, the same progeny genotype results.

(3) **The complementarity of gametes:** If two female gametes of a maternal genotype are complementary to those of the pollen pool, such combinations will produce the same outcrossing progeny genotype. For example, two gametes of maternal genotype 11/10, 11 and 10 are respectively combined with complementary male gametes from the pollen pool, 00 and 01, to generate the same progeny genotype 10/10.

(4) **Double heterozygote of a maternal plant:** This type of plant contains two different diplotypes which produce the same arrays of gametes but with different relative proportions (see Table 2).

(iii) *Likelihood and estimation*

Based on the structure of offspring genotypes in Table 1, we construct a log likelihood for parameters $\Theta=(p_{11}, p_{10}, p_{01}, p_{00}, w, r)$ as

$$\log L(\Theta) = \sum_{i\geqslant i'=0}^{1}\sum_{j\geqslant j'=0}^{1}\sum_{l\geqslant l'=0}^{1}\sum_{r\geqslant r'=0}^{1} N_{ii'jj'}^{ll'rr'}\log(P_{ii'jj'}P_{ii'jj'}^{ll'rr'}), \tag{4}$$

https://doi.org/10.1017/S0016672308009932 Published online by Cambridge University Press

Likelihood (4) is implemented with the Expectation–Maximization (EM) algorithm to obtain the maximum likelihood estimates (MLEs) of Θ.

(a) *Estimation of gamete frequencies*

In the E step, calculate the expected numbers of gametes 11, 10, 01 and 00 within each observed offspring genotype derived from a maternal genotype, expressed as

$$_k\Psi_{ii'jj'}^{ll'rr'} \quad (k=11,10,01,00). \tag{5}$$

Tables 3–6 provide the formulae for estimating the expected numbers of gametes 11, 10, 01 and 00, respectively. In the M step, estimate the gamete frequencies using

$$p_k=\frac{\sum_{i\geqslant i'=0}^{1}\sum_{j\geqslant j'=0}^{1}\sum_{l\geqslant l'=0}^{1}\sum_{r\geqslant r'=0}^{1}{}_k\Psi_{ii'jj'}^{ll'rr'}N_{ii'jj'}^{ll'rr'}}{2\sum_{i\geqslant i'=0}^{1}\sum_{j\geqslant j'=0}^{1}\sum_{l\geqslant l'=0}^{1}\sum_{r\geqslant r'=0}^{1}N_{ii'jj'}^{ll'rr'}}. \tag{6}$$

(b) *Estimation of the recombination fraction*

In the E step, calculate the expected number of r within an offspring genotype derived from a maternal plant of double heterozygote using

$$R_{10/10}^{11/11}=\frac{\bar{\phi}r(wp_{11}+\bar{w}r)}{2P_{10/10}^{11/11}},\quad R_{10/10}^{11/10}=\frac{r[w(\phi p_{11}+\bar{\phi}p_{10})+\bar{w}\bar{r}]}{2P_{10/10}^{11/11}},$$

$$R_{10/10}^{11/00}=\frac{\phi r(wp_{10}+\bar{w}r)}{2P_{10/10}^{11/00}},\quad R_{10/10}^{10/11}=\frac{r[w(\phi p_{11}+\bar{\phi}p_{01})+\bar{w}\bar{r}]}{2P_{10/10}^{10/11}},$$

$$R_{10/10}^{10/10}=\frac{r[\phi w(p_{10}+p_{01})+\bar{\phi}\bar{w}(p_{11}+p_{00})+2\bar{w}r]}{2P_{10/10}^{10/10}},$$

$$R_{10/10}^{10/00}=\frac{r[w(\phi p_{00}+\bar{\phi}p_{10})+\bar{w}\bar{r}]}{2P_{10/10}^{10/00}},\quad R_{10/10}^{00/11}=\frac{\phi r(wp_{01}+\bar{w}r)}{2P_{10/10}^{00/11}},$$

$$R_{10/10}^{00/10}=\frac{r[w(\phi p_{00}+\bar{\phi}p_{01})+\bar{w}\bar{r}]}{2P_{10/10}^{00/10}},\quad R_{10/10}^{00/00}=\frac{\bar{\phi}r(wp_{00}+\bar{w}r)}{2P_{10/10}^{00/00}}. \tag{7}$$

In the M step, estimate the recombination fraction using

$$r=\frac{\sum_{l\geqslant l'=0}^{1}\sum_{r\geqslant r'=0}^{1}R_{10/10}^{ll'rr'}N_{10/10}^{ll'rr'}}{\sum_{l\geqslant l'=0}^{1}\sum_{r\geqslant r'=0}^{1}N_{10/10}^{ll'rr'}}. \tag{8}$$

(c) *Estimation of outcrossing rate*

In the E step, calculate the expected number of w within each possible offspring genotype derived from a maternal genotype using

$$W_{ii'jj'}^{ll'rr'}=\frac{wO_{ii'jj'}^{ll'rr'}}{\bar{w}S_{ii'jj'}^{ll'rr'}+wO_{ii'jj'}^{ll'rr'}}. \tag{9}$$

In the M step, calculate the outcrossing rate using

$$w=\frac{\sum_{i\geqslant i'=0}^{1}\sum_{j\geqslant j'=0}^{1}\sum_{l\geqslant l'=0}^{1}\sum_{r\geqslant r'=0}^{1}W_{ii'jj'}^{ll'rr'}N_{ii'jj'}^{ll'rr'}}{\sum_{i\geqslant i'=0}^{1}\sum_{j\geqslant j'=0}^{1}\sum_{l\geqslant l'=0}^{1}\sum_{r\geqslant r'=0}^{1}N_{ii'jj'}^{ll'rr'}}. \tag{10}$$

Note that not all offspring genotypes contain w if they are not derived from a double heterozygote maternal plant (see Table 1), although the summation over all possible genotypes is generally given.

An iterative loop of E and M steps between equations (5), (7) and (9) and equations (6), (8) and (10) is constructed to estimate the parameters. After haplotype frequencies are estimated, allele frequencies at two markers and their LD (D) are estimated as

$$\hat{p}=\hat{p}_{11}+\hat{p}_{10},$$
$$\hat{q}=\hat{p}_{11}+\hat{p}_{01},$$
$$\hat{D}=\hat{p}_{11}\hat{p}_{00}-\hat{p}_{10}\hat{p}_{01}.$$

(iv) *Hypothesis testing*

The genetic parameters estimated, i.e. the LD (D), outcrossing rate (w) and recombination fraction (r), can be used to describe the genetic structure of a population. These parameters should be tested for their significance. The following hypotheses are formulated:

$$H_0\colon D=0 \quad vs. \quad H_1\colon D\neq 0,$$
$$H_0\colon w=0 \quad vs. \quad H_1\colon w\neq 0,$$
$$H_0\colon w=1 \quad vs. \quad H_1\colon w\neq 1,$$
$$H_0\colon r=0{\cdot}5 \quad vs. \quad H_1\colon r\neq 0{\cdot}5.$$

For each hypothesis, the likelihoods, $L_0(\tilde{\Theta})$ and $L_1(\tilde{\Theta})$, are calculated, respectively, where the tilde corresponds to the MLEs for the null hypothesis and the hat corresponds to the MLEs for the alternative hypothesis. The log-likelihood ratio test statistic is then calculated by using

$$\mathrm{LR}=-2[\ln L_0(\tilde{\Theta})-\ln L_1(\hat{\Theta})], \tag{11}$$

which is asymptotically χ^2-distributed with one degree of freedom. The estimates of the parameters under each null hypothesis should be derived separately.

3. Computer simulation

(i) *Design*

Computer simulation was conducted to examine the statistical properties of the two-locus model for estimating the LD, outcrossing rate and recombination fraction between different molecular markers in a natural population. We consider a set of OP families randomly derived from a natural population, in which the genotype distribution of two given markers were simulated with their frequencies. The frequencies of

https://doi.org/10.1017/S0016672308009932 Published online by Cambridge University Press

Table 3. *The expected number* ($_{11}\Psi^{ll'rr'}_{ii'jj'}$) *of gamete 11, within an offspring genotype derived from a maternal genotype. Note that the double heterozygote is obtained by dividing the expression in the table by both the frequency of maternal genotype* ($P_{ii'jj'}$) *and the overall frequency of the corresponding offspring genotype* ($P^{ll'rr'}_{ii'jj'}$), *whereas* $_{11}\Psi^{ll'rr'}_{ii'jj'}$ *for all the genotypes is calculated by dividing the expression only by the overall frequency of the corresponding offspring genotypes*

Maternal genotype	Offspring								
	11/11	11/10	11/00	10/11	10/10	10/00	00/11	00/10	00/00
11/11	$2\bar{w}+3wp_{11}$	$2wp_{10}$	0	$2wp_{01}$	$2wp_{00}$	0	0	0	0
11/10	$\frac{1}{4}\bar{w}+wp_{11}$	$\frac{1}{2}(\bar{w}+2wp_{11}+wp_{10})$	$\frac{1}{4}\bar{w}+\frac{1}{2}wp_{10}$	$\frac{1}{2}wp_{01}$	$\frac{1}{2}w(p_{01}+p_{00})$	$\frac{1}{2}wp_{00}$	0	0	0
11/00	0	wp_{11}	0	0	0	0	0	0	0
10/11	$\frac{1}{4}\bar{w}+wp_{11}$	$\frac{1}{2}wp_{10}$	0	$\frac{1}{2}(\bar{w}+2wp_{11}+wp_{01})$	$\frac{1}{2}w(p_{10}+p_{00})$	0	$\frac{1}{4}\bar{w}+\frac{1}{2}wp_{01}$	$\frac{1}{2}wp_{00}$	0
10/10	$p_{11}(2w\bar{r}p_{11}p_{00}+\frac{1}{2}\bar{w}\bar{r}^2p_{00}+wrp_{10}p_{01})$	$p_{11}(w\bar{r}p_{10}p_{00}+2wrp_{11}p_{00}+\bar{w}r\bar{r}p_{00}+w\bar{r}p_{10}p_{01})$	$p_{11}p_{00}r(wp_{10}+\frac{1}{2}\bar{w}r)$	$p_{11}[w\bar{r}p_{00}p_{01}+2wrp_{11}p_{00}+\bar{w}\,\bar{r}rp_{00}+w\bar{r}p_{10}p_{01}]$	$p_{11}p_{00}[w\bar{r}(p_{00}+2p_{11})+wr(p_{01}+p_{10})+\bar{w}(\bar{r}^2+r^2)]+wrp_{11}p_{10}p_{01}$	$p_{11}p_{00}(wrp_{00}+w\bar{r}p_{10}+\bar{w}\bar{r}r)$	$p_{11}p_{00}(wrp_{01}+\frac{1}{2}\bar{w}r^2)$	$p_{11}p_{00}(wrp_{00}+w\bar{r}p_{01}+\bar{w}\bar{r}r)$	$p_{11}p_{00}\bar{r}(wp_{00}+\frac{1}{2}\bar{w}\bar{r})$
10/00	0	$\frac{1}{2}wp_{11}$	0	0	$\frac{1}{2}wp_{11}$	0	0	0	0
00/11	0	0	0	wp_{11}	0	0	0	0	0
00/10	0	0	0	$\frac{1}{2}wp_{11}$	$\frac{1}{2}wp_{11}$	0	0	0	0
00/00	0	0	0	0	wp_{11}	0	0	0	0

Table 4. *The expected number* ($_{10}\Psi^{ll'rr'}_{ii'jj'}$) *of gamete 10, within an offspring genotype derived from a maternal genotype. Note that* $_{10}\Psi^{ll'rr'}_{ii'jj'}$ *for the double heterozygote is obtained by dividing the expression in the table by both the frequency of maternal genotype* ($P_{ii'jj'}$) *and the overall frequency of the corresponding offspring genotype* ($P^{ll'rr'}_{ii'jj'}$), *whereas* $_{10}\Psi^{ll'rr'}_{ii'jj'}$ *for all the genotypes is calculated by dividing the expression only by the overall frequency of the corresponding offspring genotypes*

Maternal genotype	Offspring								
	11/11	11/10	11/00	10/11	10/10	10/00	00/11	00/10	00/00
11/11	0	wp_{10}	0	0	0	0	0	0	0
11/10	$\frac{1}{4}\bar{w}+\frac{1}{2}wp_{11}$	$\frac{1}{2}(\bar{w}+2wp_{10}+wp_{11})$	$\frac{1}{4}\bar{w}+wp_{10}$	$\frac{1}{2}wp_{01}$	$\frac{1}{2}w(p_{01}+p_{00})$	$\frac{1}{2}wp_{00}$	0	0	0
11/00	0	$2wp_{11}$	$2\bar{w}+3wp_{10}$	0	$2wp_{01}$	$2wp_{00}$	0	0	0
10/11	0	$\frac{1}{2}wp_{10}$	0	0	$\frac{1}{2}wp_{10}$	0	0	0	0
10/10	$p_{11}p_{01}(wrp_{11}+\frac{1}{2}\bar{w}r^2)$	$p_{10}(w\bar{r}p_{11}p_{00}+2wrp_{01}p_{10}+w\bar{r}p_{11}p_{01}+\bar{w}\bar{r}rp_{01})$	$p_{10}(wrp_{11}p_{00}+2w\bar{r}p_{10}p_{01}+\frac{1}{2}\bar{w}\bar{r}^2)$	$p_{10}p_{01}[wrp_{01}+w\bar{r}p_{11}+\bar{w}\bar{r}r]$	$p_{10}p_{01}[wr(p_{00}+p_{11})+w\bar{r}(p_{01}+2p_{10})+\bar{w}(r^2+\bar{r}^2)]+wrp_{10}p_{11}p_{00}$	$p_{10}[w\bar{r}(p_{11}p_{00}+p_{01}p_{00})+2wrp_{10}p_{01}+\bar{w}\bar{r}rp_{01}]$	$p_{10}p_{01}(w\bar{r}p_{01}+\frac{1}{2}\bar{w}\bar{r}^2)$	$p_{10}p_{01}(w\bar{r}p_{00}+wrp_{01}+\bar{w}\bar{r}r)$	$p_{10}p_{01}r(wp_{00}+\frac{1}{2}\bar{w}r)$
10/00	0	$\frac{1}{2}wp_{11}$	$\frac{1}{4}\bar{w}+wp_{10}$	0	$\frac{1}{2}w(p_{11}+p_{01})$	$\frac{1}{2}(\bar{w}+2wp_{10}+wp_{00})$	0	$2wp_{01}$	$\frac{1}{4}\bar{w}+wp_{00}$
00/11	0	0	0	0	wp_{10}	0	0	0	0
00/10	0	0	0	0	$\frac{1}{2}wp_{10}$	$\frac{1}{2}wp_{10}$	0	0	0
00/00	0	0	0	0	0	wp_{10}	0	0	0

Table 5. *The expected number* ($_{01}\Psi^{ll'rr'}_{ii'jj'}$) *of gamete 01, within an offspring genotype derived from a maternal genotype. Note that* $_{01}\Psi^{ll'rr'}_{ii'jj'}$ *for the double heterozygote is obtained by dividing the expression in the table by both the frequency of the maternal genotype* ($P_{ii'jj'}$) *and the overall frequency of the corresponding offspring genotype* ($P^{ll'rr'}_{ii'jj'}$), *whereas* $_{01}\Psi^{ll'rr'}_{ii'jj'}$ *for all the genotypes is calculated by dividing the expression only by the overall frequency of the corresponding offspring genotypes*

Maternal genotype	Offspring								
	11/11	11/10	11/00	10/11	10/10	10/00	00/11	00/10	00/00
11/11	0	0	0	wp_{01}	0	0	0	0	0
11/10	0	0	0	$\frac{1}{2}wp_{01}$	$\frac{1}{2}wp_{01}$	0	0	0	0
11/00	0	0	0	0	wp_{01}	0	0	0	0
10/11	$\frac{1}{4}\bar{w}+wp_{11}$	$\frac{1}{2}wp_{10}$	0	$\frac{1}{2}(\bar{w}+2wp_{01}+wp_{11})$	$\frac{1}{2}w(p_{10}+p_{00})$	0	$\frac{1}{4}\bar{w}+wp_{01}$	$\frac{1}{2}wp_{00}$	0
10/10	$p_{10}p_{01}(wrp_{11}+\frac{1}{2}\bar{w}r^2)$	$p_{10}p_{01}(wrp_{10}+w\bar{r}p_{11}+\bar{w}\bar{r}r)$	$p_{10}p_{01}(w\bar{r}p_{10}+\frac{1}{2}\bar{w}\bar{r}^2)$	$p_{01}(wrp_{00}p_{11}+2wrp_{10}+\bar{w}\bar{r}rp_{10})$	$p_{10}p_{01}[wr(p_{00}+p_{11})+w\bar{r}(2p_{01}+p_{10})+\bar{w}(r^2+\bar{r}^2)]+wrp_{01}p_{11}p_{00}$	$p_{10}p_{01}(w\bar{r}p_{00}+wrp_{10}+\bar{w}\bar{r}r)$	$p_{10}p_{01}(w\bar{r}p_{00}+2w\bar{r}p_{10}p_{01}+\frac{1}{2}\bar{w}\bar{r}^2)$	$p_{01}[w\bar{r}(p_{11}p_{00}+p_{10}p_{00})+2wrp_{01}p_{10}+\bar{w}\bar{r}rp_{10}]$	$p_{10}p_{01}r(wp_{00}+\frac{1}{2}\bar{w}r)$
10/00	0	0	0	0	$\frac{1}{2}wp_{01}$	0	0	$\frac{1}{2}wp_{01}$	0
00/11	0	0	0	$2wp_{11}$	$2wp_{10}$	0	$2\bar{w}+3wp_{01}$	$2wp_{00}$	0
00/10	0	0	0	$\frac{1}{2}wp_{11}$	$\frac{1}{2}w(p_{11}+p_{10})$	$\frac{1}{2}wp_{10}$	$\frac{1}{4}\bar{w}+wp_{01}$	$\frac{1}{2}(\bar{w}+2wp_{01}+wp_{00})$	$\frac{1}{4}\bar{w}+\frac{1}{2}wp_{00}$
00/00	0	0	0	0	0	0	0	wp_{01}	0

Table 6. *The expected number* (${}_{00}\Psi^{ll'rr'}_{ii'jj'}$) *of gamete 00, within an offspring genotype derived from a maternal genotype. Note that* ${}_{00}\Psi^{ll'rr'}_{ii'jj'}$ *for the double heterozygote is obtained by dividing the expression in the table by both the frequency of maternal genotype* ($P_{ii'jj'}$) *and the overall frequency of the corresponding offspring genotype* ($P^{ll'rr'}_{ii'jj'}$), *whereas* ${}_{00}\Psi^{ll'rr'}_{ii'jj'}$ *for all the genotypes is calculated by dividing the expression only by the overall frequency of the corresponding offspring genotypes*

Maternal genotype	Offspring								
	11/11	11/10	11/00	10/11	10/10	10/00	00/11	00/10	00/00
11/11	0	0	0	0	wp_{00}	0	0	0	0
11/10	0	0	0	0	$\frac{1}{2}wp_{00}$	$\frac{1}{2}wp_{00}$	0	0	0
11/00	0	0	0	0	0	wp_{00}	0	0	0
10/11	0	0	0	0	$\frac{1}{2}wp_{00}$	0	0	$\frac{1}{2}wp_{00}$	0
10/10	$p_{11}p_{00}(w\bar{r}p_{11}+\frac{1}{2}\bar{w}\bar{r}^2)$	$p_{11}p_{00}(w\bar{r}p_{10}+wrp_{11}+\bar{w}\bar{r}r)$	$p_{11}p_{00}r(wp_{10}+\frac{1}{2}\bar{w}r)$	$p_{11}p_{00}(w\bar{r}p_{01}+wrp_{11}+\bar{w}\bar{r}r)$	$p_{11}p_{00}[w\bar{r}(p_{11}+2p_{00})+wr(p_{01}+p_{10})]+wrp_{01}p_{10}p_{00}$	$p_{00}[p_{11}(2wrp_{00}+2\bar{w}\bar{r}r)+\bar{w}(\bar{r}^2+r^2)+w\bar{r}(p_{11}p_{10}+p_{10}p_{01})]+\bar{w}\bar{r}rp_{11}$	$p_{11}p_{00}(wrp_{01}+\frac{1}{2}\bar{w}r^2)$	$p_{00}[2wrp_{11}p_{00}+w\bar{r}(p_{11}p_{01}+p_{10}p_{01})]$	$p_{00}(2w\bar{r}p_{11}p_{00}+\frac{1}{2}\bar{w}\bar{r}^2p_{11}+wrp_{10}p_{01})$
10/00	0	$\frac{1}{2}wp_{11}$	$\frac{1}{4}\bar{w}+\frac{1}{2}wp_{10}$	0	$\frac{1}{2}w(p_{11}+p_{01})$	$\frac{1}{2}(\bar{w}+2wp_{00}+p_{10})$	0	$\frac{1}{2}wp_{01}$	$\frac{1}{4}\bar{w}+wp_{00}$
00/11	0	0	0	0	0	0	0	wp_{00}	0
00/10	0	0	0	$\frac{1}{2}wp_{11}$	$\frac{1}{2}w(p_{11}+p_{10})$	$\frac{1}{2}wp_{10}$	$\frac{1}{4}\bar{w}+\frac{1}{2}wp_{01}$	$\frac{1}{2}(\bar{w}+2wp_{00}+wp_{01})$	$\frac{1}{4}\bar{w}+wp_{00}$
00/00	0	0	0	0	$2wp_{11}$	$2wp_{10}$	0	$2wp_{01}$	$2\bar{w}+3wp_{00}$

Table 7. *MLEs of parameters and their standard errors (in parentheses) obtained from 100 simulation replicates with the (small family number × large family size) sampling strategy*

	r		*w*				*D*	
No.	True	MLE	True	MLE	*p* MLE	*q* MLE	True	MLE
1	0·05	0·0503 (0·0134)	0·1	0·1012 (0·0153)	0·5994 (0·0108)	0·5005 (0·0118)	0·02	0·0200 (0·0079)
2	0·05	0·0505 (0·0114)	0·1	0·0992 (0·0157)	0·6012 (0·0105)	0·4992 (0·0119)	0·10	0·1011 (0·0066)
3	0·05	0·0466 (0·0310)	0·5	0·5014 (0·0295)	0·5989 (0·0115)	0·4986 (0·0108)	0·02	0·0195 (0·0069)
4	0·05	0·0464 (0·0218)	0·5	0·5027 (0·0267)	0·6006 (0·0108)	0·5001 (0·0094)	0·10	0·1001 (0·0049)
5	0·05	0·1494 (0·1912)	0·9	0·8966 (0·0298)	0·6019 (0·0102)	0·4997 (0·0092)	0·02	0·0211 (0·0060)
6	0·05	0·0550 (0·0510)	0·9	0·9005 (0·0308)	0·6003 (0·0101)	0·5021 (0·0099)	0·10	0·1011 (0·0053)
7	0·25	0·2561 (0·0362)	0·1	0·1000 (0·0144)	0·6003 (0·0121)	0·4987 (0·0117)	0·02	0·0202 (0·0075)
8	0·25	0·2457 (0·0277)	0·1	0·0987 (0·0146)	0·5996 (0·0104)	0·5000 (0·0107)	0·10	0·0999 (0·0063)
9	0·25	0·2527 (0·0770)	0·5	0·5093 (0·0289)	0·6003 (0·0096)	0·4999 (0·0114)	0·02	0·0203 (0·0066)
10	0·25	0·2466 (0·0449)	0·5	0·5012 (0·0313)	0·6015 (0·0098)	0·5016 (0·0097)	0·10	0·1007 (0·0059)
11	0·25	0·3056 (0·2484)	0·9	0·8986 (0·0301)	0·6007 (0·0090)	0·5005 (0·0094)	0·02	0·0200 (0·0055)
12	0·25	0·2531 (0·0585)	0·9	0·9023 (0·0276)	0·6003 (0·0097)	0·4998 (0·0110)	0·10	0·0998 (0·0055)

two-locus genotypes (derived from diplotypes) are determined by gamete frequencies, specified by allele frequencies and LD in a population and the recombination fraction for a given maternal plant. We will consider the influences of different outcrossing rates (i.e. $w = 0{\cdot}1$ for low, 0·5 for medium and 0·9 for high), different recombination fractions (i.e. $r = 0{\cdot}05$ for a strong linkage and 0·25 for a weak linkage) and different linkage disequilibria (i.e. $D = 0{\cdot}02$ for strong independence and 0·10 for weak independence) on parameter estimation. We will consider all these possible combinations of parameter values.

To provide practical guidance on the use of this model, we simulate marker data with three different sampling strategies. A fixed number of samples (say 1000) can be allocated among and within OP families. We will use three sampling strategies: (1) small family number × large family size' (10×100), (2) moderate family number × moderate family size (32×32), and large family number × small family size (100×10). Results under each of these strategies will be given.

(ii) *Results*

Tables 7–9 summarize the simulation results with different parameters and sampling strategies. In general, the model provides reasonable estimates of all parameters, although the accuracy and precision of parameter estimates depend on the values of these parameters, sampling strategies and interactions among all the factors. The estimation of the recombination fraction tends to prefer the 'small family number × large family size' sampling strategy (Table 7). It appears that the 'large family number × small family size' sampling strategy is favourable for the estimation of population genetic parameters including allele frequencies, LD and outcrossing rate (Table 9). The 'moderate family number × moderate family size' sampling strategy is somewhat in between (Table 8). In all the strategies, the estimation precision of allele frequencies and LD increases with increasing outcrossing rate. The recombination fraction can be estimated more precisely when the two markers are strongly linked. It is interesting to see that increasing LD leads to better estimation of the recombination fraction. As expected, increasing the outcrossing rate reduces the estimation precision of the recombination fraction. Especially, when outcrossing rate is very high ($w = 0{\cdot}9$), the recombination fraction will be poorly estimated for the two markers that are strongly independent.

There is reasonably good power for detecting a significant LD and linkage between two markers, although such a power varies with the values of the parameters (results not shown). It seems that the power detection is not sensitive to sampling strategies. There are marked interactions in the power sensitivity between parameter values. The power of linkage detection decreases with increasing outcrossing rate, whereas the power of LD detection increases with increasing outcrossing rate.

4. Discussion

The past two decades have witnessed a dramatic increase of interest in molecular marker technologies and their applications to study the genetic structure of a natural population and map QTLs (quantitative trait loci) responsible for a quantitative trait (Reich *et al*., 2001; Ardlie *et al*., 2002; Dawson *et al*., 2002; Gabriel *et al*., 2002; reviewed in Georges 2007). In this paper, we have proposed an algorithmic approach for constructing the linkage–linkage disequilibrium map of a genome by genotyping a set of OP seeds sampled from a natural population. By estimating several key

https://doi.org/10.1017/S0016672308009932 Published online by Cambridge University Press

Table 8. *MLEs of parameters and their standard errors* (*in parentheses*) *obtained from 100 simulation replicates with the* (*moderate family number* × *moderate family size*) *sampling strategy*

	r		*w*		*p*	*q*	*D*	
No.	True	MLE	True	MLE	MLE	MLE	True	MLE
1	0·05	0·0498 (0·0141)	0·1	0·0992 (0·0149)	0·6010 (0·0118)	0·5024 (0·0116)	0·02	0·0195 (0·0080)
2	0·05	0·0497 (0·0119)	0·1	0·0998 (0·0175)	0·6015 (0·0119)	0·5012 (0·0109)	0·10	0·0993 (0·0070)
3	0·05	0·0473 (0·0317)	0·5	0·4999 (0·0295)	0·5999 (0·0112)	0·4995 (0·0101)	0·02	0·0194 (0·0070)
4	0·05	0·0510 (0·0235)	0·5	0·5004 (0·0262)	0·5989 (0·0091)	0·4997 (0·0110)	0·10	0·0992 (0·0053)
5	0·05	0·1339 (0·2014)	0·9	0·8995 (0·0289)	0·5984 (0·0085)	0·5011 (0·0089)	0·02	0·0210 (0·0057)
6	0·05	0·0595 (0·0486)	0·9	0·8981 (0·0274)	0·6004 (0·0095)	0·5000 (0·0095)	0·10	0·1002 (0·0053)
7	0·25	0·2507 (0·0347)	0·1	0·0993 (0·0126)	0·6001 (0·0104)	0·5004 (0·0108)	0·02	0·0215 (0·0082)
8	0·25	0·2497 (0·0270)	0·1	0·0972 (0·0166)	0·6013 (0·0112)	0·5017 (0·0102)	0·10	0·1006 (0·0063)
9	0·25	0·2746 (0·0886)	0·5	0·4960 (0·0272)	0·6003 (0·0097)	0·5013 (0·0102)	0·02	0·0196 (0·0072)
10	0·25	0·2577 (0·0368)	0·5	0·5011 (0·0301)	0·5991 (0·0106)	0·5008 (0·0126)	0·10	0·0997 (0·0063)
11	0·25	0·2735 (0·2497)	0·9	0·8969 (0·0281)	0·6004 (0·0105)	0·4985 (0·0090)	0·02	0·0204 (0·0064)
12	0·25	0·2553 (0·0612)	0·9	0·8991 (0·0307)	0·6003 (0·0093)	0·4993 (0·0083)	0·10	0·1002 (0·0055)

Table 9. *MLEs of parameters and their standard errors* (*in parentheses*) *obtained from 100 simulation replicates with the* (*large family number* × *small family size*) *sampling strategy*

	r		*w*		*p*	*q*	*D*	
No.	True	MLE	True	MLE	MLE	MLE	True	MLE
1	0·05	0·0508 (0·0145)	0·1	0·0984 (0·0148)	0·5993 (0·0102)	0·4979 (0·0106)	0·02	0·0206 (0·0066)
2	0·05	0·0517 (0·0113)	0·1	0·1007 (0·0165)	0·5983 (0·0109)	0·4981 (0·0105)	0·10	0·1007 (0·0062)
3	0·05	0·0556 (0·0344)	0·5	0·5007 (0·0232)	0·6007 (0·0124)	0·5009 (0·0128)	0·02	0·0199 (0·0068)
4	0·05	0·0508 (0·0231)	0·5	0·5008 (0·0279)	0·5977 (0·0112)	0·4985 (0·0119)	0·10	0·1002 (0·0065)
5	0·05	0·1274 (0·2088)	0·9	0·9039 (0·0271)	0·6004 (0·0094)	0·5002 (0·0091)	0·02	0·0217 (0·0058)
6	0·05	0·0530 (0·0510)	0·9	0·9042 (0·0274)	0·5997 (0·0093)	0·5011 (0·0092)	0·10	0·1004 (0·0046)
7	0·25	0·2602 (0·0590)	0·1	0·0996 (0·0140)	0·6000 (0·0127)	0·4995 (0·0116)	0·02	0·0195 (0·0079)
8	0·25	0·2532 (0·0298)	0·1	0·0996 (0·0139)	0·5996 (0·0110)	0·4991 (0·0092)	0·10	0·0997 (0·0060)
9	0·25	0·2581 (0·0857)	0·5	0·4979 (0·0308)	0·6014 (0·0086)	0·4991 (0·0088)	0·02	0·0210 (0·0077)
10	0·25	0·2456 (0·0374)	0·5	0·4977 (0·0279)	0·6000 (0·0101)	0·4987 (0·0098)	0·10	0·0987 (0·0057)
11	0·25	0·3293 (0·2558)	0·9	0·8975 (0·0302)	0·5989 (0·0091)	0·5012 (0·0107)	0·02	0·0202 (0·0063)
12	0·25	0·2556 (0·0701)	0·9	0·9000 (0·0294)	0·5997 (0·0090)	0·4992 (0·0097)	0·10	0·1013 (0·0057)

population genetic parameters, i.e. the relative occurrence of selfing and outcrossing, LD, heterozygosity (estimated from allele frequencies) and recombination fraction, this approach will provide a tool for better understanding the pattern and organization of genetic variation in outcrossing populations. Furthermore, by elucidating the relationship between the linkage and LD in terms of the so-called LD map, the new algorithm can be used to infer the evolutionary history and process of natural populations and to identify genes for disease or yield traits (Remington *et al*., 2001; Ardlie *et al*., 2002; Farnir *et al*., 2000; McRae *et al*., 2002; Rafalski & Morgante, 2004).

The new approach capitalizes on the outcrossing nature of plants, allowing a certain proportion of selfing. Outcrossing is a common characteristic of many plants, including economically and ecologically important species like poplar, eucalyptus, pine and spruce (Butcher & Southerton, 2007; Miller & Schaal, 2006). This approach will find its immediate application in the genetic research of these important but understudied species. It has three significant advantages. First, it is simple and easily deployed in practice. By sampling and genotyping half-sib seeds from multiple maternal plants in a population, the approach provides the estimation of important population genetic parameters. Second, we derived a group of EM-based closed forms for parameter estimation for the OP-based sampling strategy, greatly facilitating the computing process of the parameters. The accuracy and precision of parameter estimation are affected by many factors including sample size and parameter range. A reasonable sampling strategy including the relative importance of family number and family size can be readily determined from simulation studies. Third, the approach allows the test of a number of meaningful hypotheses about the linkage, LD and outcrossing rate, providing a quantitative framework for understanding the genetic structure of a natural outcrossing population.

https://doi.org/10.1017/S0016672308009932 Published online by Cambridge University Press

The approach can be extended to consider multiallelic markers and dominant markers. Unlike many annual crops, forest trees are still in wild or semi-wild conditions, in which there is a rich source of variation due to many alleles at a single gene. Multiallelic markers like microsatellites are a vital tool for the population genetic study of forest trees. On the other hand, for many underrepresented organisms, some economically cheap dominant markers are still useful although their informativeness is limited (Kuang *et al*., 1998; Silbiger *et al*., 1998; Kremer *et al*., 2005). When the OP-based sampling strategy considers multiallelic or dominant markers, new, more complicated algorithms need to be derived. Li *et al*. (2007) derived a model for the LD between dominant markers in a diploid population. Their model can be integrated with our OP-based sampling strategy to provide a comprehensive estimation of population genetic parameters with dominant markers. The computer code of the proposed algorithm is available from the corresponding author upon request.

This work was partially supported by Joint NSF/NIH grant number DMS/NIGMS-0540745 and NNSFC grant number 30771752.

References

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

Butcher, P. A. & Southerton, S. (2007). Marker-assisted selection in forestry species. In *Marker-assisted Selection: Current Status and Future Perspectives in Crops, Livestock, Forestry and Fish* (ed. E. Guimaraes, J. Ruane, B. Scherf, A. Sonnino and J. Dargie), pp. 283–305, chapter 15. Rome: FAO.

Charlesworth, D. (2003). The effects of inbreeding on the genetic diversity of populations. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences* **358**, 1051–1070.

Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lohmmussaar, E., Zernant, J., Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D. R., Cardon, L. R. & Dunham, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548.

Farnir, F., Coppieters, W., Arranz, J.-J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Moreau, L., Mni, M., Nezer, C., Simon, P., Vanmanshoven, P., Wagenaar, D. & Georges, M. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Research* **10**, 220–227.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

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.

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

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

Ingvarsson, P. K. (2002). A metapopulation perspective of genetic diversity and differentiation in partially self-fertilizing plants. *Evolution* **56**, 2368–2373.

Ingvarsson, P. K. (2005). Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics* **169**, 945–953.

Kremer, A., Caron, H., Cavers, S., Colpaert, N., Gheysen, G., Gribe, R., Lemes, M., Lowe, A. J., Margis, R., Navarro, C. & Salgueiro, F. (2005). Monitoring genetic diversity in tropical trees with multilocus dominant markers. *Heredity* **95**, 274–280.

Kuang, H., Richardson, T. E., Carson, S. D. & Bongarten, B. C. (1998). An allele responsible for seedling death in *Pinus radiata* D. Don. *Theoretical and Applied Genetics* **96**, 640–644.

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

Li, Y. C., Li, Y., Wu, S., Han, K., Wang, Z. J., Hou, W., Zeng, Y. R. & Wu, R. L. (2007). Estimation of linkage disequilibria in diploid populations with multilocus dominant markers. *Genetics* **176**, 1879–1892.

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

McRae, A. F., McEwan, J. C., Dodds, K. G., Wilson, T., Crawford, A. M. & Slate, J. (2002). Linkage disequilibrium in domestic sheep. *Genetics* **160**, 1113–1122.

Miller, A. J. & Schaal, B. A. (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.

Nordborg, M. (2000). Linkage disequilibrium, gene trees and selfing: ancestral recombination graph with partial self-fertilization. *Genetics* **154**, 923–929.

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

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

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M. & Buckler, E. S. IV (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the USA* **98**, 11479–11484.

Silbiger, R. N., Christ, S. A., Leonard, A. C., Garg, M., Lattier, D. L., Dawes, S., Dimsoski, P., McCormick, F., Wessendarp, T., Gordon, D. A., Roth, A. C., Smith, M. K. & Toth, G. P. (1998). Preliminary studies on the population genetics of the central stoneroller (*Campostoma anomalum*) from the Great Miami River Basin, Ohio. *Environmental Monitoring and Assessment* **51**, 481–495.

Tishkoff, S. A. & Williams, S. M. (2002). Genetic analysis of African populations: human evolution and complex disease. *Nature Reviews Genetics* **3**, 611–621.

Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., Krings, M., Pääbo, S., Watson, E., Risch, N., Jenkins, T. & Kidd, K. K. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387.

Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., Piro, A., Stoneking, M., Tagarelli, A., Tagarelli, G., Touma, E. H., Williams, S. M. & Clark, A. G. (2001). Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**, 455–462.

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

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