A Multivalent Pairing Model of Linkage Analysis in Autotetraploids

Samuel S. Wu,*,[1] Rongling Wu,*,[1] Chang-Xing Ma,*,† Zhao-Bang Zeng,‡ Mark C. K. Yang* and George Casella*

**Department of Statistics, University of Florida, Gainesville, Florida 32611, †Department of Statistics, Nankai University, Tianjin 300071, People's Republic of China and ‡Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695*

Manuscript received May 2, 2001
Accepted for publication August 29, 2001

ABSTRACT

Polyploidy has been recognized as an important step in the evolutionary diversification of flowering plants and may have a significant impact on plant breeding. Statistical analyses for linkage mapping in polyploid species can be difficult due to considerable complexities in polysomic inheritance. In this article, we develop a novel statistical method for linkage analysis of polymorphic markers in a full-sib family of autotetraploids. This method is established on multivalent pairings of homologous chromosomes at meiosis and can provide a simultaneous maximum-likelihood estimation of the double reduction frequencies of and recombination fraction between two markers. The EM algorithm is implemented to provide a tractable way for estimating relative proportions of different modes of gamete formation that generate identical gamete genotypes due to multivalent pairings. Extensive simulation studies were performed to demonstrate the statistical properties of this method. The implications of the new method for understanding the genome structure and organization of polyploid species are discussed.

POLYPLOIDY is an important evolutionary force in flowering plants (Stebbins 1971; Grant 1981; Bever and Felber 1992; Jackson and Jackson 1996; Soltis and Soltis 2000). It is estimated that as much as 30–80% of angiosperms are polyploids or have experienced one or more episodes of polyploidization (Stebbins 1971; Grant 1981; Masterson 1994). Evidence for the creative role of polyploidy in evolution is well synthesized in a recent review by Otto and Whitton (2000), although they estimated that only 2–4% of speciation events in flowering plants involve polyploidization. The frequency of polyploidy in domesticated plant taxa is also high (75%); alfalfa, banana, canola, coffee, cotton, potato, soybean, strawberry, sugarcane, sweet potato, and wheat represent excellent examples of polyploids of economic importance (Hilu 1993). To study the evolutionary consequences of polyploidy on genome organization and develop superior varieties of polyploid plant species, a number of genome projects have now been launched to construct genetic linkage maps using molecular markers and identify genes responsible for economically important traits in polyploid populations ranging from tetraploid (potato) to octoploid (sugarcane; Wu *et al.* 1992; da Silva *et al.* 1993; Yu and Pauls 1993; Grivet *et al.* 1996; Hackett *et al.* 1998; Meyer *et al.* 1998; Brouwer and Osborn 1999; Ripol *et al.* 1999).

For allopolyploids derived from the chromosome combination of distinct genomes and subsequent chromosome doubling (Soltis and Soltis 2000), statistical methods developed for molecular linkage mapping by estimating recombination fractions between different loci in diploid species (Lander and Green 1987) will also apply. However, these methods cannot be used in autopolyploids that are formed due to the chromosome doubling of the same genome by fusion of unreduced gametes (Soltis and Soltis 2000). Autopolyploids may undergo either bivalent (two chromosomes pair) or multivalent pairing (more than two chromosomes pair) or both, at meiosis, in which a gene has more than one possible partner (or set of partners). Polysomic inheritance could result from the multivalent formation. Most of the available statistical methods for autopolyploid linkage analysis assume bivalent pairings (Wu *et al.* 1992; Hackett *et al.* 1998; Ripol *et al.* 1999; Luo *et al.* 2000, 2001). Statistical analysis assuming multivalent pairings has not been explored thoroughly because of the complexity of polysomic inheritance.

Double reduction is a phenomenon that two sister chromatids of a chromosome sort into the same gamete (Darlington 1929; de Winton and Haldane 1931; Mather 1936; Fisher 1947). It may be generated due to multivalent pairings in autopolyploids. Figure 1 shows how different types of gametes are formed. At anaphase I, chromatids located on a chromosome may migrate either to the same pole (reductional separation) or to different poles (equational separation). The type of separation depends on the number and the type of crossovers located between the centromere and the locus under consideration. We consider the segregation of

Corresponding author: Samuel S. Wu, Division of Biostatistics, P. O. Box 100212, University of Florida, Gainesville, FL 32610. E-mail: samwu@biostat.ufl.edu

[1] These authors contributed equally to this work.

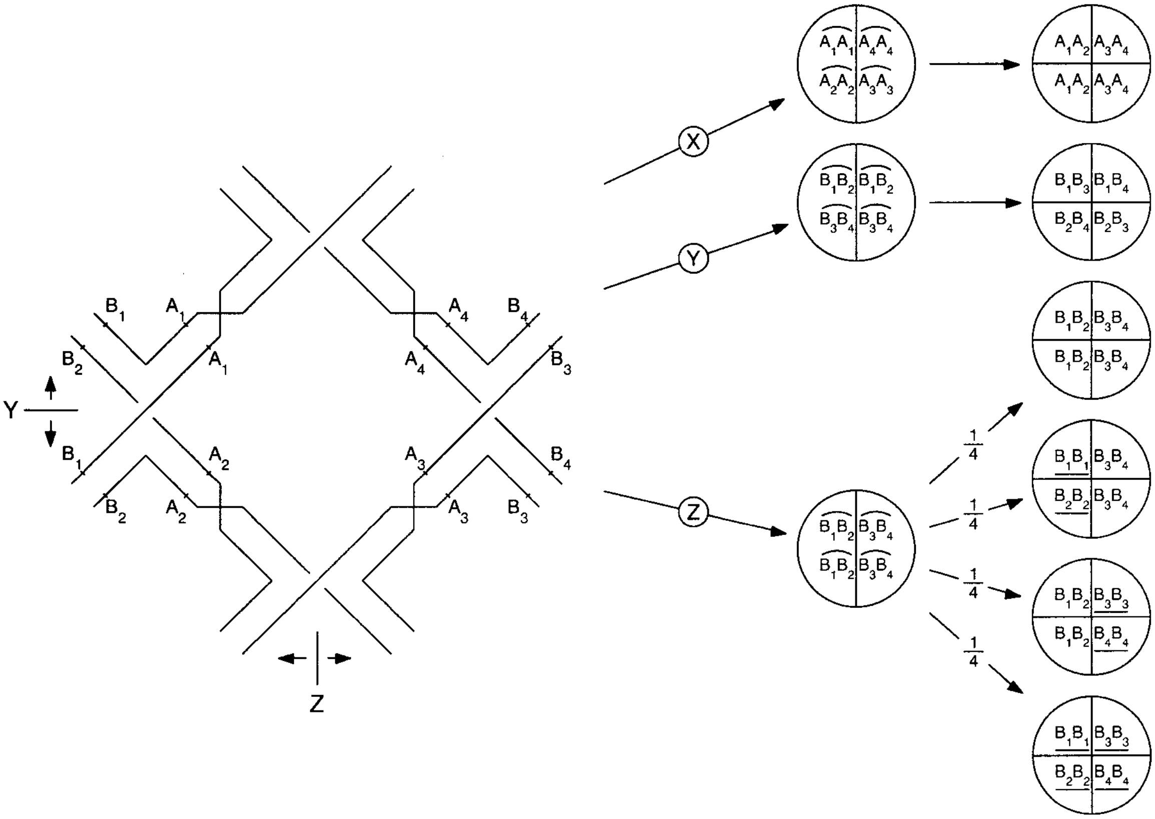

Figure 1.—A diagram displaying the segregation patterns of loci A and B during meiosis in an autopolyploid (modified from Mather 1936 and Bever and Felber 1992). Locus A having no crossover with the centromere undergoes path X of reductional separation (no double reduction), whereas locus B displaying a crossover with the centromere undergoes either path Y of equational separation with no double reduction or path Z of equational separation with double reduction. Gametes having undergone double reductions are underscored.

two loci A and B in autotetraploid demonstrating quadrivalent formation during meiosis. Locus A is so close to the centromere that no crossover happened between them. The first division for this locus is reductional and double reduction never occurs (path X). Locus B has one crossover with the centromere and, thus, undergoes equational separation. If the four homologous chromosomes segregate randomly, they may migrate to the same cell in two different ways. In the first way, chromosomes 1 and 2 and their respective homologues migrate to the same cells, and therefore alleles located on sister chromatids reach different gametes and double reduction never occurs (path Y). In the second way, chromosomes 1 and 2 and chromosomes 3 and 4 migrate to the same cells, which may cause double reduction when chromatids segregate randomly.

Fisher (1947) formulated a pioneering theoretical model for analyzing two linked loci in an autotetraploid undergoing quadrivalent pairings during meiosis. Although Fisher elegantly described the modes of gamete formation in terms of the recombination number between the two loci and the frequency of double reduction at each locus, he was not able to provide a tractable computational method for estimating these parameters. In this article, we use Fisher's model to devise a maximum-likelihood method for simultaneously estimating the frequency of double reduction and the recombination fraction between different markers in autopolyploids whose gamete formation is predominately due to multivalent pairings. The method relied on an expectation-maximization (EM) algorithm (Dempster *et al.* 1977). Mathematically, we prove that the difference in the frequency of double reduction between two loci is bounded by two times the recombination fraction in tetraploid. Our linkage analysis here is based on fully informative codominant markers of eight different alleles at each marker between the two autotetraploid parents. Statistical properties of this autopolyploid method are examined using a simulation study.

AUTOTETRAPLOID MODEL

A quadrivalent pairing model for two linked markers: Consider two linked markers $\mathcal{M}^k$ and $\mathcal{M}^l$ on the same chromosome in an autotetraploid. At marker $\mathcal{M}^k$, four alleles, each assigned to one of the four homologous chromosomes, are labeled by P_1^k, P_2^k, P_3^k, and P_4^k for parent P and by Q_1^k, Q_2^k, Q_3^k, and Q_4^k for parent Q. Accordingly, four different alleles at marker $\mathcal{M}^l$ are labeled by P_1^l, P_2^l, P_3^l, and P_4^l for parent P and by Q_1^l, Q_2^l, Q_3^l, and Q_4^l for parent Q. The recombination fraction between the two markers is denoted by θ_P for parent P and θ_Q for parent Q. For the two autotetraploid parents used for the cross, there are a total of 576 allelic configurations or linkage phase assignments between the two markers, one of which is schematically expressed as

$$\begin{array}{c|c|c|c|} P_1^k & P_2^k & P_3^k & P_4^k \\ P_1^l & P_2^l & P_3^l & P_4^l \end{array} \times \begin{array}{c|c|c|c|} Q_1^k & Q_2^k & Q_3^k & Q_4^k \\ Q_1^l & Q_2^l & Q_3^l & Q_4^l \end{array}, \tag{1}$$

where lines indicate the individual homologous chromosomes on which the two markers are located. The recombination fractions θ_P and θ_Q are estimated on the basis of the segregation of the two-marker joint genotypes observed in the progeny of the family. However, the observations of the joint marker genotypes are confounded by the models of meiotic pairings (bivalent or quadrivalent) and parental linkage phases of different alleles across the two maternally and two paternally derived chromosomes. To make accurate estimates for θ_P and θ_Q, therefore, it is essential to select a most likely pairing model and linkage phase configuration over the two parents.

In this article, we proposed a model for fully informative codominant markers, *i.e.*, those of eight different alleles between the two autotetraploid parents at each marker. We assume that the four homologous chromosomes form quadrivalents. Thus, for a particular marker $\mathcal{M}^k$, we must consider the full chromatid complement that may be represented as gametes $P_1^kP_1^k$, $P_2^kP_2^k$, $P_3^kP_3^k$, and $P_4^kP_4^k$ for parent P and gametes $Q_1^kQ_1^k$, $Q_2^kQ_2^k$, $Q_3^kQ_3^k$, and $Q_4^kQ_4^k$ for parent Q. The generation of these gametes is typical of the four-strand model in which both chromatids of a single chromosome may be passed to the same gamete, forming the so-called double reduction (Darlington 1929; Mather 1936; Fisher 1947). The frequency of double reduction is a constant for any given locus, depending on its distance from the centromere. We denote the frequencies of double reduction at $\mathcal{M}^k$ by α_P for parent P and α_Q for parent Q. Similarly, β_P and β_Q are denoted for marker $\mathcal{M}^l$. Following the classification in Fisher (1947), with two linked loci in tetrasomics there are four different combinations in terms of the existence of double reduction:

1. Both markers display double reductions;
2. only marker $\mathcal{M}^k$ displays double reductions;
3. only marker $\mathcal{M}^l$ displays double reductions; and
4. none of the markers display double reductions.

Since there are four sources for the allele at any given locus, a gametic chromosome with two loci can be made up 16 ways:

$$\begin{array}{c|} P_r^k \\ P_s^l \end{array} \; (r,\, s = 1,\, 2,\, 3,\, 4).$$

Each gamete has two chromosomes and these will be of $\frac{1}{2} \times 16 \times 17 = 136$ different possible types. For one parent, all these types of gametes can be classified into 11 basic modes according to double reduction and the number of recombination events (Mather 1936; Table 1):

1 and 2: The first two modes of gametic formation shown in Table 1 involve double reduction at both markers $\mathcal{M}^k$ and $\mathcal{M}^l$. Of the two modes, the first has no recombination between the chromosomes and thus entails only four parental types of gamete:

$$\begin{array}{c|c|} P_1^k & P_1^k \\ P_1^l & P_1^l \end{array}, \begin{array}{c|c|} P_2^k & P_2^k \\ P_2^l & P_2^l \end{array}, \begin{array}{c|c|} P_3^k & P_3^k \\ P_3^l & P_3^l \end{array}, \begin{array}{c|c|} P_4^k & P_4^k \\ P_4^l & P_4^l \end{array}.$$

The second mode has 12 possibilities as a result of recombination between a pair of the four chromosomes:

$$\begin{array}{c|c|} P_1^k & P_1^k \\ P_2^l & P_2^l \end{array}, \begin{array}{c|c|} P_1^k & P_1^k \\ P_3^l & P_3^l \end{array}, \begin{array}{c|c|} P_1^k & P_1^k \\ P_4^l & P_4^l \end{array}; \begin{array}{c|c|} P_2^k & P_2^k \\ P_1^l & P_1^l \end{array}, \begin{array}{c|c|} P_2^k & P_2^k \\ P_3^l & P_3^l \end{array}, \begin{array}{c|c|} P_2^k & P_2^k \\ P_4^l & P_4^l \end{array};$$

$$\begin{array}{c|c|} P_3^k & P_3^k \\ P_1^l & P_1^l \end{array}, \begin{array}{c|c|} P_3^k & P_3^k \\ P_2^l & P_2^l \end{array}, \begin{array}{c|c|} P_3^k & P_3^k \\ P_4^l & P_4^l \end{array}; \begin{array}{c|c|} P_4^k & P_4^k \\ P_1^l & P_1^l \end{array}, \begin{array}{c|c|} P_4^k & P_4^k \\ P_2^l & P_2^l \end{array}, \begin{array}{c|c|} P_4^k & P_4^k \\ P_3^l & P_3^l \end{array}.$$

3 and 4: The second two modes include double reduction only at marker $\mathcal{M}^k$. For mode 3, one parental chromosome is unchanged, but the other is made up by all possible types of recombination between this chromosome and the remaining three. There are 12 possibilities and typical gametes are like

$$\begin{array}{c|c|} P_1^k & P_1^k \\ P_1^l & P_2^l \end{array}, \begin{array}{c|c|} P_1^k & P_1^k \\ P_1^l & P_3^l \end{array}, \begin{array}{c|c|} P_1^k & P_1^k \\ P_1^l & P_4^l \end{array}.$$

And for mode 4, both chromosomes are derived from recombination between the four parental chromosomes. There are also 12 possibilities such as

$$\begin{array}{c|c|} P_1^k & P_1^k \\ P_2^l & P_3^l \end{array}, \begin{array}{c|c|} P_1^k & P_1^k \\ P_2^l & P_4^l \end{array}, \begin{array}{c|c|} P_1^k & P_1^k \\ P_3^l & P_4^l \end{array}.$$

5 and 6: The next two modes involve double reduction only at marker $\mathcal{M}^l$ and have classifications similar to the second two modes.

Other 5: The last five modes (7A, 7B, 8A, 8B, and 9), in which neither marker $\mathcal{M}^k$ nor $\mathcal{M}^l$ has double reduction, can be sorted into three types. In the first type, mode 7, two gametic chromosomes are derived from two of the parental chromosomes either without recombination (mode 7A) or with recombination (mode 7B). There are six possibilities for each group. Typical gamete types are

$$\begin{array}{c|c|} P_1^k & P_2^k \\ P_1^l & P_2^l \end{array} \quad \text{and} \quad \begin{array}{c|c|} P_1^k & P_2^k \\ P_2^l & P_1^l \end{array}.$$

Because the same genotype is represented, 7A and 7B *cannot be distinguished* on the basis of the marker

TABLE 1

The model of quadrivalent formation for a diploid gamete in a tetrasomic parent with two linked markers $\mathcal{M}^k$ and $\mathcal{M}^l$

Double reduction combination	Mode	Typical gamete	No. of types	Relative frequency of formation	No. of recombination events
Both markers	1	$\begin{array}{c\|c\|} P_1^k & P_1^k \\ P_1^l & P_1^l \end{array}$	4	f_1	0
	2	$\begin{array}{c\|c\|} P_1^k & P_1^k \\ P_2^l & P_2^l \end{array}$	12	f_2	2
Only marker $\mathcal{M}^k$	3	$\begin{array}{c\|c\|} P_1^k & P_1^k \\ P_1^l & P_2^l \end{array}$	12	f_3	1
	4	$\begin{array}{c\|c\|} P_1^k & P_1^k \\ P_2^l & P_3^l \end{array}$	12	f_4	2
Only marker $\mathcal{M}^l$	5	$\begin{array}{c\|c\|} P_1^k & P_2^k \\ P_1^l & P_1^l \end{array}$	12	f_5	1
	6	$\begin{array}{c\|c\|} P_2^k & P_3^k \\ P_1^l & P_1^l \end{array}$	12	f_6	2
Neither marker	7A	$\begin{array}{c\|c\|} P_1^k & P_2^k \\ P_1^l & P_2^l \end{array}$	6	$f_{7\mathrm{A}}$	0
	7B	$\begin{array}{c\|c\|} P_1^k & P_2^k \\ P_2^l & P_1^l \end{array}$	6	$f_{7\mathrm{B}}$	2
	8A	$\begin{array}{c\|c\|} P_1^k & P_2^k \\ P_1^l & P_3^l \end{array}$	24	$f_{8\mathrm{A}}$	1
	8B	$\begin{array}{c\|c\|} P_1^k & P_2^k \\ P_3^l & P_1^l \end{array}$	24	$f_{8\mathrm{B}}$	2
	9	$\begin{array}{c\|c\|} P_1^k & P_2^k \\ P_3^l & P_4^l \end{array}$	12	f_9	2
Total			136	1	

There are 136 different gametes with the two linked markers

$$\begin{array}{c|c|} P_{r_1}^k & P_{r_2}^k \\ P_{s_1}^l & P_{s_2}^l \end{array} \quad (r_1,\ r_2,\ s_1,\ s_2 = 1,\ 2,\ 3,\ 4).$$

We use a single f_7 to denote the mixed frequency with which both 7A and 7B occur at meiosis; *i.e.*, $f_7 = f_{7\mathrm{A}} + f_{7\mathrm{B}}$. For the same reason, the mixed frequency of 8A and 8B is denoted by f_8.

phenotypes. The second type (mode 8) of nondouble reduction is that two gametic chromosomes are derived from three of the parental chromosomes with one event of recombination (8A, 24 possibilities) or two events of recombinations (8B, 24 possibilities). Gamete examples for modes 8A and 8B are

$$\begin{array}{c|c|} P_1^k & P_2^k \\ P_1^l & P_3^l \end{array} \quad \text{and} \quad \begin{array}{c|c|} P_1^k & P_2^k \\ P_3^l & P_1^l \end{array}.$$

They are *also indistinguishable* because they have identical genotypes. The third type (mode 9) of nondouble reduction includes recombination between all four different chromosomes such as

$$\begin{array}{c|c|} P_1^k & P_2^k \\ P_3^l & P_4^l \end{array}.$$

Mode 9 has 12 possibilities.

Because gametes for fully informative markers are unique to the two parents and because the two parents are assumed to behave independently in terms of double reduction and recombination, gamete genotypes can provide adequate information for linkage analysis as much as zygote genotypes. Therefore, to simplify our treatments, we base our linkage analysis on the segregation of the gamete genotypes in each parent. Thereafter, only parent P is considered because a symmetrical inference can be made for parent Q. We refer to the frequencies of double reduction and recombination fraction between the markers for parent P by α, β, and θ without the subscript P, unless otherwise specified.

Parameter estimation: For marker $\mathcal{M}^k$, assume a fixed assignment for the four alleles of parent P in the order P_1^k, P_2^k, P_3^k, and P_4^k. Given such a fixed assignment for marker $\mathcal{M}^k$, we randomly assign the four observed alleles of marker $\mathcal{M}^l$, P_1^l, P_2^l, P_3^l, and P_4^l, with a total of 24 different possibilities. One of the possibilities should present a correct assignment for the alleles of the two markers among the four homologous chromosomes. The estimates of the frequencies of double reduction and the recombination fraction between the two markers should be based on their best, but unknown, allelic assignment across the parental chromosomes. For linkage analysis in autotetraploid populations, therefore, a vector of unknown parameters can be denoted by $\boldsymbol{\pi}_\omega = (A_\omega, \alpha, \beta, \theta)^T$, where A_ω is the ωth allelic assignment for marker $\mathcal{M}^l$ relative to the fixed allelic assignment of marker $\mathcal{M}^k$.

Given a particular allelic assignment for parent P as shown in expression (1), four double reduction gametes and six nondouble reduction gametes generated by marker $\mathcal{M}^k$ can be arrayed in the order $\{P^k_1P^k_1, P^k_2P^k_2, P^k_3P^k_3, P^k_4P^k_4, P^k_1P^k_2, P^k_1P^k_3, P^k_1P^k_4, P^k_2P^k_3, P^k_2P^k_4, P^k_3P^k_4\}$ and $\{P^l_1P^l_1, P^l_2P^l_2, P^l_3P^l_3, P^l_4\ P^l_4, P^l_1\ P^l_2, P^l_1P^l_3, P^l_1P^l_4, P^l_2P^l_3, P^l_2P^l_4, P^l_3P^l_4\}$ at marker $\mathcal{M}^l$. Thus, we can identify $10 \times 10 = 100$ two-marker gamete genotypes for parent P. Following notation in Fisher (1947), we define f, the relative frequencies of the 11 different modes of gamete formation, which must sum to unity (Table 1). However, because the marker phenotypes of 7A and 7B cannot be distinguished, we use a single f_7 to denote the mixed frequency with which both 7A and 7B occur at meiosis. For the same reason, the mixed frequency of 8A and 8B is denoted by f_8. It is not difficult to express the joint relative frequencies of two-marker diploid gametes in matrix notation:

$$\mathbf{H} = \{H^{r_1r_2}_{s_1s_2}\}_{10\times10}$$

$$\begin{array}{r|cccccccccc|}
 & P^k_1P^k_1 & P^k_2P^k_2 & P^k_3P^k_3 & P^k_4P^k_4 & P^k_1P^k_2 & P^k_1P^k_3 & P^k_1P^k_4 & P^k_2P^k_3 & P^k_2P^k_4 & P^k_3P^k_4 \\
P^l_1P^l_1 & \frac{1}{4}f_1 & \frac{1}{12}f_2 & \frac{1}{12}f_2 & \frac{1}{12}f_2 & \frac{1}{12}f_5 & \frac{1}{12}f_5 & \frac{1}{12}f_5 & \frac{1}{12}f_6 & \frac{1}{12}f_6 & \frac{1}{12}f_6 \\
P^l_2P^l_2 & \frac{1}{12}f_2 & \frac{1}{4}f_1 & \frac{1}{12}f_2 & \frac{1}{12}f_2 & \frac{1}{12}f_5 & \frac{1}{12}f_6 & \frac{1}{12}f_6 & \frac{1}{12}f_5 & \frac{1}{12}f_5 & \frac{1}{12}f_6 \\
P^l_3P^l_3 & \frac{1}{12}f_2 & \frac{1}{12}f_2 & \frac{1}{4}f_1 & \frac{1}{12}f_2 & \frac{1}{12}f_6 & \frac{1}{12}f_5 & \frac{1}{12}f_6 & \frac{1}{12}f_5 & \frac{1}{12}f_6 & \frac{1}{12}f_5 \\
P^l_4P^l_4 & \frac{1}{12}f_2 & \frac{1}{12}f_2 & \frac{1}{12}f_2 & \frac{1}{4}f_1 & \frac{1}{12}f_6 & \frac{1}{12}f_6 & \frac{1}{12}f_5 & \frac{1}{12}f_6 & \frac{1}{12}f_5 & \frac{1}{12}f_5 \\
= P^l_1P^l_2/P^l_2P^l_1 & \frac{1}{12}f_3 & \frac{1}{12}f_3 & \frac{1}{12}f_4 & \frac{1}{12}f_4 & \frac{1}{6}f_7 & \frac{1}{24}f_8 & \frac{1}{24}f_8 & \frac{1}{24}f_8 & \frac{1}{24}f_8 & \frac{1}{6}f_9 \\
P^l_1P^l_3/P^l_3P^l_1 & \frac{1}{12}f_3 & \frac{1}{12}f_4 & \frac{1}{12}f_3 & \frac{1}{12}f_4 & \frac{1}{24}f_8 & \frac{1}{6}f_7 & \frac{1}{24}f_8 & \frac{1}{24}f_8 & \frac{1}{6}f_9 & \frac{1}{24}f_8 \\
P^l_1P^l_4/P^l_4P^l_1 & \frac{1}{12}f_3 & \frac{1}{12}f_4 & \frac{1}{12}f_4 & \frac{1}{12}f_3 & \frac{1}{24}f_8 & \frac{1}{24}f_8 & \frac{1}{6}f_7 & \frac{1}{6}f_9 & \frac{1}{24}f_8 & \frac{1}{24}f_8 \\
P^l_2P^l_3/P^l_3P^l_2 & \frac{1}{12}f_4 & \frac{1}{12}f_3 & \frac{1}{12}f_3 & \frac{1}{12}f_4 & \frac{1}{24}f_8 & \frac{1}{24}f_8 & \frac{1}{6}f_9 & \frac{1}{6}f_7 & \frac{1}{24}f_8 & \frac{1}{24}f_8 \\
P^l_2P^l_4/P^l_4P^l_2 & \frac{1}{12}f_4 & \frac{1}{12}f_3 & \frac{1}{12}f_4 & \frac{1}{12}f_3 & \frac{1}{24}f_8 & \frac{1}{6}f_9 & \frac{1}{24}f_8 & \frac{1}{24}f_8 & \frac{1}{6}f_7 & \frac{1}{24}f_8 \\
P^l_3P^l_4/P^l_4P^l_3 & \frac{1}{12}f_4 & \frac{1}{12}f_4 & \frac{1}{12}f_3 & \frac{1}{12}f_3 & \frac{1}{6}f_9 & \frac{1}{24}f_8 & \frac{1}{24}f_8 & \frac{1}{24}f_8 & \frac{1}{24}f_8 & \frac{1}{6}f_7
\end{array}. \quad (2)$$

However, as illustrated earlier, there are as many as 136 gamete formations for any two linked markers. The 36 "extra" gamete formations are each due to a reciprocal allelic assignment of marker $\mathcal{M}^l$ and are located in the $6 \times 6 = 36$ cells of the above matrix's bottom-right corner, in which neither of the two markers displays double reduction (Table 1). Of these 36 formations, 6 are under mode 7, 24 are under mode 8, and the remaining 6 are under mode 9. For example, gamete formations

$$\begin{array}{c|c|} P^k_1 & P^k_2 \\ P^l_1 & P^l_2 \end{array} \quad \text{and} \quad \begin{array}{c|c|} P^k_1 & P^k_2 \\ P^l_2 & P^l_1 \end{array}$$

are two reciprocal assignments, but they have the same genotype and are mixed in the same cell at row 5 and column 5.

Because formation mode 7 is a mixture of double recombinants and nonrecombinants, the determination of the expected number of recombination events under this mode requires information about the relative proportions of these two types of offspring. Given the relative proportion of double recombinants in mode 7 ($\phi = f_{7B}/f_7$, see appendix), the expected number of recombination events is 2ϕ. Similarly, for mode 8, which is a mixture of single recombinants and double recombinants, the expected number of recombination events is calculated as $1 \cdot (1 - \psi) + 2 \cdot \psi = 1 + \psi$, where ψ is the proportion of double recombinants in mode 8 ($\psi = f_{8B}/f_8$; see appendix). The expected numbers of recombination events between the two markers can be expressed in matrix notation as

$$\mathbf{D} = \{D^{r_1r_2}_{s_1s_2}\}_{10\times10}$$

$$\begin{array}{r|cccccccccc|}
 & P^k_1P^k_1 & P^k_2P^k_2 & P^k_3P^k_3 & P^k_4P^k_4 & P^k_1P^k_2 & P^k_1P^k_3 & P^k_1P^k_4 & P^k_2P^k_3 & P^k_2P^k_4 & P^k_3P^k_4 \\
P^l_1P^l_1 & 0 & 2 & 2 & 2 & 1 & 1 & 1 & 2 & 2 & 2 \\
P^l_2P^l_2 & 2 & 0 & 2 & 2 & 1 & 2 & 2 & 1 & 1 & 2 \\
P^l_3P^l_3 & 2 & 2 & 0 & 2 & 2 & 1 & 2 & 1 & 2 & 1 \\
P^l_4P^l_4 & 2 & 2 & 2 & 0 & 2 & 2 & 1 & 2 & 1 & 1 \\
= P^l_1P^l_2/P^l_2P^l_1 & 1 & 1 & 2 & 2 & 2\phi & 1+\psi & 1+\psi & 1+\psi & 1+\psi & 2 \\
P^l_1P^l_3/P^l_3P^l_1 & 1 & 2 & 1 & 2 & 1+\psi & 2\phi & 1+\psi & 1+\psi & 2 & 1+\psi \\
P^l_1P^l_4/P^l_4P^l_1 & 1 & 2 & 2 & 1 & 1+\psi & 1+\psi & 2\phi & 2 & 1+\psi & 1+\psi \\
P^l_2P^l_3/P^l_3P^l_2 & 2 & 1 & 1 & 2 & 1+\psi & 1+\psi & 2 & 2\phi & 1+\psi & 1+\psi \\
P^l_2P^l_4/P^l_4P^l_2 & 2 & 1 & 2 & 1 & 1+\psi & 2 & 1+\psi & 1+\psi & 2\phi & 1+\psi \\
P^l_3P^l_4/P^l_4P^l_3 & 2 & 2 & 1 & 1 & 2 & 1+\psi & 1+\psi & 1+\psi & 1+\psi & 2\phi
\end{array}.$$

The above information allows us to express the recombination fraction θ and the two double reduction parameters, α at marker $\mathcal{M}^k$ and β at marker $\mathcal{M}^l$, in terms of $f_1, \ldots, f_9$ and ϕ, ψ. We have

$$\alpha = f_1 + f_2 + f_3 + f_4;$$

$$\beta = f_1 + f_2 + f_5 + f_6;$$

$$2\theta = (f_3 + f_5) + 2(f_2 + f_4 + f_6 + f_9) + 2\phi f_7 + (1 + \psi)f_8.$$

From the above equations, it follows that $|\alpha - \beta| = |f_3 + f_4 - f_5 - f_6) \leq f_3 + f_4 + f_5 + f_6 \leq 2\theta$. Therefore the difference in the frequency of double reduction between two loci is bounded by two times the recombination fraction in tetraploid. This inequality is consistent with the fact that when two markers are close, their double reduction rates tend to be similar. We believe similar inequalities exist for other ploidy levels. However, due to complexity of gamete types for those cases, we are not able to generalize the result at this moment.

For a fully informative marker, every gamete genotype can be well distinguished. Thus, N offspring in a full-sib family can be sorted into the nine distinguishable gamete formation modes of size $N_1, N_2, \ldots, N_9$, respectively (see Table 1). It is not difficult to derive the explicit expressions of the maximum-likelihood estimates for the frequencies of these nine formation modes $f_1, f_2, \ldots, f_9$ in terms of the corresponding sample frequencies $N_1, N_2, \ldots, N_9$ on the basis of the following likelihood function given the observed marker data (**M**):

$$\ell(\boldsymbol{f}|\mathbf{M}) = \binom{N}{N_1 \ldots N_9} \prod_{i=1}^{9} f_i^{N_i}.$$

From the above matrix **H**, which indicates where double reduction has occurred for each of the markers, the two double-reduction parameters, α and β, can be estimated in terms of the corresponding frequencies of formation modes; *i.e.*, $\hat{\alpha} = (N_1 + N_2 + N_3 + N_4)/N$ and $\hat{\beta} = (N_1 + N_2 + N_5 + N_6)/N$. Since these are simply estimates of binomial proportions, the variances of $\hat{\alpha}$ and $\hat{\beta}$ are $\alpha(1 - \alpha)/N$ and $\beta(1 - \beta)/N$, respectively.

Suppose we could distinguish the two f_7 modes and the two f_8 modes; the likelihood function given complete data ($N_1, N_2, \ldots, N_6, N_{7A}, N_{7B}, N_{8A}, N_{8B}, N_9$) is

$$\ell(\boldsymbol{f}|\mathbf{M}) = \binom{N}{N_1 \ldots N_9} \left(\prod_{i=1}^{6} f_i^{N_i}\right) f_{7A}^{N_{7A}} f_{7B}^{N_{7B}} f_{8A}^{N_{8A}} f_{8B}^{N_{8B}} f_9^{N_9}. \quad (3)$$

On the basis of the observed incomplete data $N_1, N_2, \ldots, N_7, N_8, N_9$, the EM algorithm is used to estimate the recombination fraction by maximizing the likelihood Equation 3 (Dempster *et al.* 1977; Lander and Green 1987). The general equations formulating the iteration of the $\tau + 1$)th EM step are given as follows:

E step: Calculate the expected number of recombination events between markers $\mathcal{M}^k$ and $\mathcal{M}^l$ for all offspring with no occurrence of double reduction. This is equivalent to estimating ϕ for mode 7 and ψ for mode 8, respectively, by

$$\hat{\phi}^{(\tau)} = \frac{[\theta^{(\tau)}]^2}{2[\theta^{(\tau)}]^2 - 18\theta^{(\tau)} + 9}, \quad \hat{\psi}^{(\tau)} = \frac{\theta^{(\tau)}}{3 - 2\theta^{(\tau)}}. \quad (4)$$

M step: Maximize the expected log-likelihood of θ. This gives an updated estimate for the recombination fraction and is obtained as

$$\hat{\theta}^{(\tau+1)} = \frac{1}{2N}[N_3 + N_5 + 2(N_2 + N_4 + N_6 + N_9) + 2\hat{\phi}^{(\tau)}N_7 + (1 + \hat{\psi}^{(\tau)})N_8]. \quad (5)$$

These two steps are repeated until the estimate of θ converges to a stable value. Such a stable value is the maximum-likelihood estimate (MLE) of θ.

If we plug ϕ and ψ from Equations 4 into 5, we can see that the stable values of the iterative procedure are solutions of the following polynomial equation in θ:

$$\theta = \frac{1}{2N}[N_3 + N_5 + 2(N_2 + N_4 + N_6 + N_9) + \frac{2\theta^2}{2\theta^2 - 18\theta + 9}N_7 + \frac{3 - \theta}{3 - 2\theta}N_8]. \quad (6)$$

Since this a fourth-order polynomial of θ, closed-form solutions exist and can be calculated very easily.

The characterization of linkage phase: We derived statistical procedures for estimating α, β, and θ when the allelic assignment as shown in expression (1) is assumed. The estimates of parameters (α, β, θ) for any one of the other 23 assignments can be similarly obtained by changing the positions of the corresponding elements in matrices **H** and **D**. One remaining issue is how to determine the best assignment, *i.e.*, one corresponding to a most likely parental linkage phase of the two markers. The most likely linkage phase can be determined using the posterior probability of $\boldsymbol{\pi}_\omega = (A_\omega, \alpha, \beta, \theta)^T$ conditional on the marker data **M**, where A_ω is the ωth allelic assignment for marker $\mathcal{M}^l$ relative to the fixed allelic assignment of marker $\mathcal{M}^k$. From Bayes' theorem:

$$P(\boldsymbol{\pi}_\omega|\mathbf{M}) = \frac{P(\boldsymbol{\pi}_\omega)P(\mathbf{M}|\boldsymbol{\pi}_\omega)}{\Sigma_{\omega=1}^{24} P(\boldsymbol{\pi}_\omega)P(\mathbf{M}|\boldsymbol{\pi}_\omega)}.$$

These posterior probabilities for all possible assignments depend on the prior probabilities $P(\boldsymbol{\pi}_\omega)$. In practice, the prior distribution can be assumed to be uniform among all 24 assignments and, in this case, the posterior probabilities are proportional to the likelihoods $L(\boldsymbol{\pi}_\omega) = P(\mathbf{M}|\boldsymbol{\pi}_\omega)$. The final MLEs for the parameters (α, β, θ) are based on the most likely assignment with the highest posterior probability.

Sved (1964) demonstrated that, unless they solely form bivalents, autotetraploids have a recombination fraction bounded by $1 - 1/x$, where x is the level of ploidy. Thus, for autotetraploids undergoing quadrivalent pairings, the maximum value of recombination fraction is $\theta = 0.75$. The test of whether or not the two given markers are linked is based on the log-likelihood-ratio test statistic under the full model (Equation 3), which corresponds to the parameter estimators derived from the most likely assignments, and the reduced model with the restraint of $\theta = 0.75$. The likelihood-ratio test (LRT) statistic calculated in this way has a χ^2-distribution with ½ d.f. under the null hypothesis (Self and Liang 1987). Thus, two markers $\mathcal{M}^k$ and $\mathcal{M}^l$ can be declared to be linked if the LRT is $> \chi^2_{1/2,\delta}$ for an appropriate choice of the type I error rate δ (for example, $\chi^2_{1/2,0.05} = 2.42$).

SIMULATION

Analysis of a simulated data set: We illustrate the autotetraploid model through analyzing a simulated example. Since gamete genotypes can provide adequate information for linkage analysis as much as zygote types, we consider analysis only on the segregation of the gamete genotypes in parent P, which is assumed to have frequencies of double reduction (0.05, 0.1) and recombination fraction 0.05. These parameters correspond to relative frequencies of the nine different gamete formation modes $\boldsymbol{f} = (0.04071, 0.00130, 0.00446, 0.00353, 0.04301, 0.01498, 0.88221, 0.00736, 0.00245)$, which give the joint relative frequencies of two-marker diploid gametes in the matrix **H.** A random sample of $N = 200$ gametes was simulated from multinomial distribution with probabilities given by **H.** The marker data **M**, *e.g.*, the counts of all gamete types, can be presented in the following matrix form:

$$\begin{array}{r} \\ P_1^lP_1^l \\ P_2^lP_2^l \\ P_3^lP_3^l \\ P_4^lP_4^l \\ \\ P_1^lP_2^l/P_2^lP_1^l \\ P_1^lP_3^l/P_3^lP_1^l \\ P_1^lP_4^l/P_4^lP_1^l \\ P_2^lP_3^l/P_3^lP_2^l \\ P_2^lP_4^l/P_4^lP_2^l \\ P_3^lP_4^l/P_4^lP_3^l \end{array} \begin{array}{c} \begin{array}{cccccccccc} P_1^kP_1^k & P_2^kP_2^k & P_3^kP_3^k & P_4^kP_4^k & P_1^kP_2^k & P_1^kP_3^k & P_1^kP_4^k & P_2^kP_3^k & P_2^kP_4^k & P_3^kP_4^k \end{array} \\ \begin{bmatrix} 3 & 0 & 0 & 0 & 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 2 & 0 \\ & & & & & & & & & \\ 0 & 2 & 0 & 0 & 22 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 16 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 33 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 35 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 36 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 31 \end{bmatrix} \end{array}.$$

Suppose parent P has alignment

$$\begin{array}{c|c|c|c|} P_1^k & P_2^k & P_3^k & P_4^k \\ P_1^l & P_2^l & P_3^l & P_4^l \end{array};$$

then there are 11 offspring in the first gamete formation mode ($N_1 = 3 + 4 + 3 + 1$). Similarly, counts for the other eight modes are $N_2 = 0$, $N_3 = 3$, $N_4 = 0$, $N_5 = 9$, $N_6 = 1$, $N_7 = 173$, $N_8 = 2$, and $N_9 = 1$. Hence we have MLEs of the relative frequencies of the nine different gamete formation modes $\hat{f} = (11/200, 0, 3/200, 0, 9/200, 1/200, 173/200, 2/200, 1/200)$, which correspond to $\hat{\alpha} = (N_1 + N_2 + N_3 + N_4)/N = 0.07$, $\hat{\beta} = (N_1 + N_2 + N_5 + N_6)/N = 0.105$, and $\hat{\theta} = 0.0453$ with log-likelihood $(\mathrm{ll})_{\mathrm{A}} = -482.98$. Furthermore, under the null hypothesis $\theta = 0.75$, the MLEs of mode frequencies are $\hat{f} = (0.01396, 0, 0.00757, 0, 0.02272, 0.25448, 0.43679, 0.01, 0.25448)$, and the parameter estimates are $\hat{\alpha} = 0.022$, $\hat{\beta} = 0.291$ with $\mathrm{ll}_{\mathrm{N}} = -616.62$.

For a second assignment

$$\begin{array}{c|c|c|c|} P_1^k & P_2^k & P_3^k & P_4^k \\ P_1^l & P_2^l & \mathbf{P_4^l} & \mathbf{P_3^l} \end{array},$$

gamete classification is different. For example, gamete

$$\begin{array}{c|c|} P_3^k & P_3^k \\ P_4^l & P_4^l \end{array}$$

has no recombination and should be classified into mode 1 instead of mode 2, and gamete

$$\begin{array}{c|c|} P_1^k & P_3^k \\ P_1^l & P_4^l \end{array}$$

should be in mode 7 instead of mode 8. The counts for all nine gamete formation modes, under the new assignments, are $N_1 = 7$, $N_2 = 7$, $N_3 = 2$, $N_4 = 1$, $N_5 = 5$, $N_6 = 5$, $N_7 = 53$, $N_8 = 123$, $N_9 = 0$. Consequently, we can obtain the MLE $(\hat{\alpha}, \hat{\beta}, \hat{\theta}) = (0.07, 0.105, 0.46436)$ with log-likelihood $\mathrm{ll}_{\mathrm{A}} = -758.50$. Similar to the first assignment, we also have MLE $(\hat{\alpha}, \hat{\beta}, \hat{\theta}) = (0.118, 0.208, 0.75)$ with log-likelihood $\mathrm{ll}_{\mathrm{N}} = -786.47$ under the null hypothesis.

This procedure needs to be repeated for all of the other 22 assignments. In Table 2, we present MLEs and log-likelihood for all 24 different allelic assignments. Figure 2 (top left) plotted the log-likelihood values against the 24 assignments of marker $\mathcal{M}^l$ with a dictionary order 1234, 1243, . . . , 4321. Estimates of the recombination fraction for different assignments are indicated by different insets in the figure. It shows that a true assignment has the largest log-likelihood value.

Since assignment 1 has the largest log-likelihood, we choose the final MLEs for the parameters on the basis of the first assignment; *e.g.*, $(\hat{\alpha}, \hat{\beta}, \hat{\theta}) = (0.07, 0.105, 0.453)$ with log-likelihood $\mathrm{ll}_{\mathrm{A}} = -482.98$. However, under the null hypothesis, the final MLE comes from the last assignment with log-likelihood $\mathrm{ll}_{\mathrm{N}} = -616.28$. Thus the LRT statistic equals $-2 \times (-616.28 + 482.98) = 266.60$, which is much larger than the cut point value $\chi^2_{1/2,0.05} = 2.42$, implying that there is very strong evidence that the two markers are linked.

More simulations: Extensive simulation studies were performed to investigate the properties of our statistical method by evaluating the effectiveness of determining a correct allelic assignment, the precision of the parameter estimates, and the power to detect linkage. A number of genetic scenarios are designed to explore the effects of different parameter values on their estimation from this new method. A segregating full-sib family of size $N = 80$, 200, 400, or 800 is simulated by hypothesizing different recombination fractions ranging from tight linkage to free recombination, $\theta = 0.05, 0.15, 0.25, 0.50, 0.65$, and 0.75, and different pairs of double reduction rates with various degrees of difference between two markers, $(\alpha, \beta) = (0.05, 0.1)$, $(0.15, 0.2)$, $(0.25, 0.3)$, $(0.1, 0.2)$ and $(0.05, 0.3)$. For $\theta = 0.05$, however, only the first three pairs of (α, β) are considered because the other two combinations are impossible (recall $|\alpha - \beta| \leq 2\theta$). The simulation is repeated 1000 times for each scenario. For each replication, the maximum-likelihood estimates $(\hat{\theta}, \hat{\alpha}, \hat{\beta})$ and the log-likelihood value are obtained for all 24 possible assignments. In addition, the LRT was calculated for each simulation to test for the significance of linkage.

In Figure 2, the log-likelihood values are plotted against the 24 different allelic assignments of marker $\mathcal{M}^l$ with a dictionary order 1234, 1243, . . . , 4321. For different assignments, different estimates of the recom-

TABLE 2

Maximum-likelihood estimates $\hat{\theta}$ of the recombination fraction with all 24 different allelic assignments of marker $\mathcal{M}^l$ for the simulated example

Allelic assignments of marker $\mathcal{M}^l$	MLE $\hat{\theta}$	Log-likelihood (ll_A) (alternative)	Log-likelihood (ll_N) (null hypothesis)
$P_1^lP_2^lP_3^lP_4^l$	0.045	−482.99	−616.62
$P_1^lP_2^lP_4^lP_3^l$	0.464	−758.50	−786.47
$P_1^lP_3^lP_2^lP_4^l$	0.417	−751.34	−784.50
$P_1^lP_3^lP_4^lP_2^l$	0.750	−730.61	−730.61
$P_1^lP_4^lP_2^lP_3^l$	0.747	−732.71	−732.72
$P_1^lP_4^lP_3^lP_2^l$	0.475	−764.19	−788.37
$P_2^lP_1^lP_3^lP_4^l$	0.480	−759.69	−783.91
$P_2^lP_1^lP_4^lP_3^l$	0.750	−600.35	−617.11
$P_2^lP_3^lP_1^lP_4^l$	0.750	−729.83	−729.83
$P_2^lP_3^lP_4^lP_1^l$	0.750	−776.58	−776.58
$P_2^lP_4^lP_1^lP_3^l$	0.750	−770.97	−770.97
$P_2^lP_4^lP_3^lP_1^l$	0.750	−729.98	−729.98
$P_3^lP_1^lP_2^lP_4^l$	0.750	−729.90	−729.90
$P_3^lP_1^lP_4^lP_2^l$	0.750	−772.89	−772.89
$P_3^lP_2^lP_1^lP_4^l$	0.467	−761.65	−790.96
$P_3^lP_2^lP_4^lP_1^l$	0.735	−732.37	−732.61
$P_3^lP_4^lP_1^lP_2^l$	0.750	−617.53	−617.53
$P_3^lP_4^lP_2^lP_1^l$	0.750	−776.98	−776.98
$P_4^lP_1^lP_2^lP_3^l$	0.750	−783.22	−783.22
$P_4^lP_1^lP_3^lP_2^l$	0.750	−730.68	−730.68
$P_4^lP_2^lP_1^lP_3^l$	0.741	−731.55	−731.63
$P_4^lP_2^lP_3^lP_1^l$	0.410	−751.53	−784.55
$P_4^lP_3^lP_1^lP_2^l$	0.750	−776.25	−776.25
$P_4^lP_3^lP_2^lP_1^l$	0.750	−616.28	−616.28

The last two columns are the log-likelihood (ll_A) under alternative hypothesis and (ll_N) under null hypothesis ($\theta = 0.75$).

bination fraction are obtained, as indicated by different insets in the figure. It is shown that a true assignment usually corresponds to the largest log-likelihood value. There is a distinct difference between the largest and the second-largest log-likelihood values, especially when θ is small. This implies that our method can well be used to characterize the marker linkage phase in parents. In some cases, the second-largest log-likelihood value is associated with the estimate of $\theta > 0.75$, so it is easy to avoid the assignment corresponding to such an estimate.

We did not report simulation results about double reduction rate estimates $\hat{\alpha}$ and $\hat{\beta}$ because we have closed-form formulas for their variances. To evaluate the precision of the recombination fraction estimates, square-rooted mean square errors (RMSEs) are calculated for all simulation scenarios (Table 3). As expected, the RMSEs decrease with increasing sample sizes. However, sample size effects also decrease with increasing sample sizes. This means that a sample size of 200–400 is adequate for providing a precise estimate of θ. It is also worth noting that the estimate works reasonably well when $N = 80$. In addition, the RMSEs of $\hat{\theta}$ values increase with decreasing θ but decrease at $\theta = 0.75$ because of the boundary effect. It is seen that the precision of $\hat{\theta}$ depends on true double reduction rates (α, β) with two tendencies (Table 3). First, the RMSEs tend to be larger when there are larger double reduction rates. Second, the RMSEs tend to increase when the difference of double reduction between the two markers increases. For example, the RMSEs of $\hat{\theta} = 0.5$ or above are larger for $(\alpha, \beta) = (0.10, 0.20)$ than $(0.25, 0.30)$, although the latter combination has larger double reduction rates.

The power to detect a significant linkage is examined on the basis of 1000 replicates (Figure 3). Obviously, the power of the test increases with increasing sample sizes. However, the effect of sample size depends on the double reduction rates and recombination fraction. For example, the effect is larger for $(\alpha, \beta) = (0.1, 0.2)$ than for $(\alpha, \beta) = (0.15, 0.2)$ when $\theta = 0.65$, but this is reversed for $\theta = 0.5$.

DISCUSSION

The main difficulty in performing linkage analysis for autopolyploids stems from the complexities of polysomic inheritance. With the occurrence of polysomic inheritance, the recombination fraction alone is no longer sufficient to specify the frequencies of gamete genotypes and their segregation patterns. To simplify linkage anal-

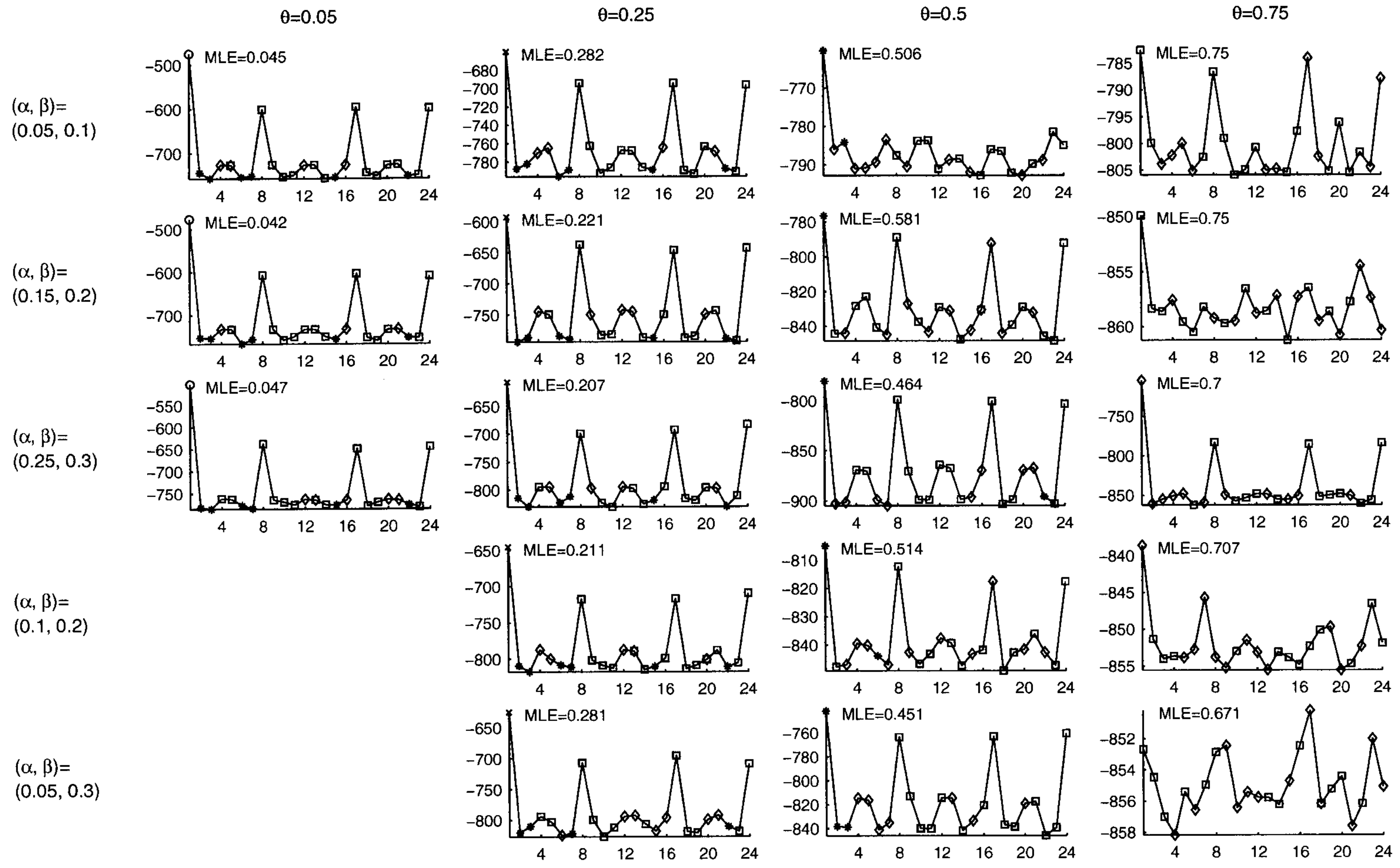

Figure 2.—Plot of the log-likelihood (ll) *vs.* 24 assignments for 18 sets of parameters (θ, α, β) from one simulation with sample size N = 200. The x-axes are the 24 assignments in dictionary order, *i.e.*, 1234, 1243, . . . , 4321. Different symbols used in the plots indicate the range of MLE $\hat{\theta}$ for each assignment: circles for $\hat{\theta}$ in (0, 1/8]; crosses for (1/8, 3/8]; stars for (3/8, 5/8]; diamonds for (5/8, 3/4]; and squares for others. The MLE $\hat{\theta}$'s corresponding to the most likely assignment are indicated in each plot.

TABLE 3

Square-rooted mean square error (RMSE) for the estimator $\hat{\theta}$ of the recombination fraction for all 28 combinations of parameters (θ, α, β) and four sample sizes *N*

		(α, β)				
θ	*N*	(0.05, 0.10)	(0.10, 0.20)	(0.15, 0.20)	(0.05, 0.30)	(0.25, 0.30)
0.05	80	0.0209	0.0236			0.0222
	200	0.0132	0.0144			0.0142
	400	0.0090	0.0101			0.0100
	800	0.0067	0.0074			0.0073
0.15	80	0.0408	0.0281	0.0375	0.0370	0.0364
	200	0.0265	0.0180	0.0245	0.0242	0.0300
	400	0.0183	0.0129	0.0173	0.0167	0.0161
	800	0.0130	0.0087	0.0118	0.0121	0.0112
0.25	80	0.0439	0.0465	0.0471	0.0487	0.0491
	200	0.0301	0.0297	0.0292	0.0321	0.0311
	400	0.0203	0.0203	0.0211	0.0226	0.0219
	800	0.0138	0.0145	0.0149	0.0153	0.0157
0.50	80	0.0793	0.1051	0.1220	0.0904	0.1076
	200	0.0396	0.0482	0.0508	0.0481	0.0471
	400	0.0267	0.0291	0.0304	0.0331	0.0296
	800	0.0201	0.0217	0.0206	0.0234	0.0197
0.65	80	0.0853	0.0733	0.0685	0.0879	0.0535
	200	0.0726	0.0492	0.0462	0.0669	0.0366
	400	0.0573	0.0351	0.0340	0.0554	0.0259
	800	0.0361	0.0258	0.0243	0.0414	0.0187
0.75	80	0.0435	0.0634	0.0528	0.0664	0.0341
	200	0.0223	0.0329	0.0230	0.0402	0.0227
	400	0.0142	0.0219	0.0143	0.0285	0.0153
	800	0.0098	0.0126	0.0107	0.0221	0.0115

ysis in autopolyploids, many earlier methods assume a pure bivalent pairing model between homologous chromosomes during meiosis (Wu *et al.* 1992; Hackett *et al.* 1998; Ripol *et al.* 1999; Luo *et al.* 2001). Although the statistical merits of these methods were demonstrated by extensive simulations, their underlying assumption may significantly deviate from biological reality. For an autopolyploid, multivalent pairings during gametogenesis may result in double reduction (Darlington 1929; de Winton and Haldane 1931; Mather 1936; Fisher 1947), a phenomenon that adds extra complexity in the establishment of a workable model for polysomic linkage analysis.

In this article, we derive a statistical method for simultaneously estimating the linkage and linkage phase between different markers in a full-sib family of autotetraploids undergoing quadrivalent pairings at meiosis. This method based on quadrivalent pairings is not a simple extension of the existing models on bivalent pairing. Rather, the method has incorporated the cytological mechanisms underlying gamete formation derived from multivalent pairings, some of which (*i.e.*, double reduction) are unique and do not happen with bivalent pairings. We also showed that the difference in the frequency of double reduction between two markers is bounded by two times their recombination fraction in tetraploid.

With these underpinning mechanisms of quadrivalent pairings, Fisher (1947) formulated a pioneering genetic model to count all possible modes of gamete formation in autotetraploids. But, in his time, he could not separate and further estimate two different modes generating the same gamete genotypes (*e.g.*, mode 7A *vs.* 7B or mode 8A *vs.* 8B; Table 1). Thanks to the development of the maximum-likelihood method implemented with the EM algorithm (Dempster *et al.* 1977), we are now able to well discriminate and estimate the proportions of these different modes by viewing them as a missing data problem.

The advantage of the EM algorithm is that it resulted in closed-form solution for the recombination fraction. However, if we forego this, it is also possible to perform a Bayesian analysis. We may assign a Dirichlet prior for the frequencies of the nine formation modes $\boldsymbol{f} = (f_1, f_2, \ldots, f_9)$, which yields a Dirichlet posterior distribution of $\boldsymbol{f}$ given the sample frequencies $N_1, N_2, \ldots, N_9$. Thus we can easily sample from the posterior of $\boldsymbol{f}$ and obtain a posterior sample of (α, β, θ) by letting $\alpha = f_1 + f_2 + f_3 + f_4$, $\beta = f_1 + f_2 + f_5 + f_6$ and solving θ using Equation 6 with each N_i/N replaced by f_i. Moreover if we extend

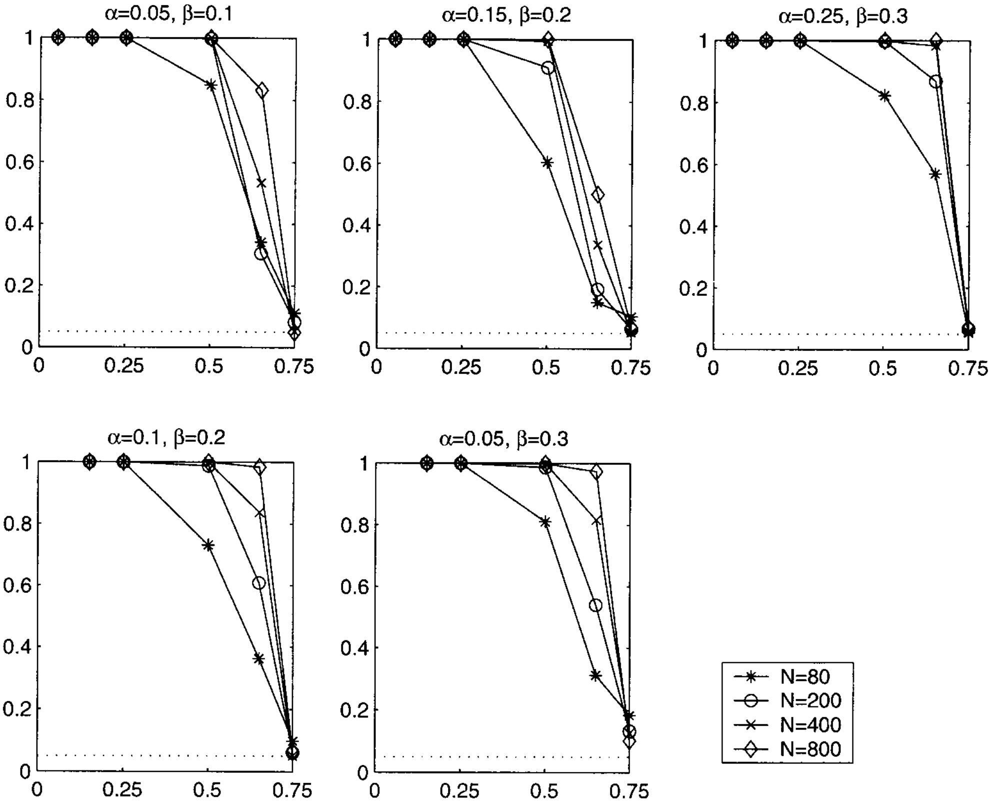

Figure 3.—The size and power of the likelihood-ratio test of linkage using $\chi^2_{1/2,0.05} = 2.42$ on the basis of 1000 replicates. The power (or size) of the test *vs.* true θ for all five sets of double reduction rates was plotted.

this to the 11 basic gamete modes $f^* = (f_1, f_2, \ldots, f_6, f_{7A}, f_{7B}, f_{8A}, f_{8B}, f_9)$, then a Gibbs sampler could be set up to obtain posterior samples (Robert and Casella 2000).

Although we have devised a statistical method for resolving a fundamentally important problem in autopolyploid linkage analysis, one that has puzzled geneticists for over one-half century, there is still much room for improvement. First, our model is proposed for fully informative codominant markers, *i.e.*, those of eight different alleles between the two autotetraploid parents at each marker. For these markers, an explicit expression exists for the MLE of the frequency of double reduction, although the estimate of the recombination fraction must rely upon EM iterations. In a practical full-sib mapping population, other types of markers, such as dominant or partially informative, may be common. For autopolyploids, dominant markers derived from randomly amplified polymorphic DNA or amplified fragment length polymorphism technologies typically cannot be distinguished among simplex (single dose), duplex (double dose), and multiplex (multiple dose) types, because they present an identical genotype (Wu *et al.* 1992; Yu and Pauls 1993; Luo *et al.* 2000). For these dominant or partially informative markers, gametes formed with double reduction may have the same genotypes as those formed without double reduction. Thus, estimating the frequency of double reduction will have to require the EM algorithm. Also, linkage analysis for these markers must be based on the segregation of zygote genotypes, because the segregation at the gamete level cannot provide adequate information for linkage analysis.

Second, our method is based on a single pairing model—quadrivalent. Chromosome pairings in autopolyploids indeed are a function of the homology between the genomes involved, with a propensity in pairing between homologous over homeologous chromosomes, which is defined as the preferential pairing factor (Sybenga 1994). Such a preferential pairing factor determines the relative importance of bivalent *vs.* multivalent pairings in autopolyploids and, therefore, can be used to model the frequency of double reduction and recombination fraction when both bivalent and multivalent pairings happen simultaneously during meiosis. Last, our method is developed for autotetraploids, but its extension to autohexaploid, autooctoploid, and autodexaploid species is important because many important plant species have such high ploidy levels (Soltis and Soltis 2000). For an autohexaploid plant, for instance, triploid gametes are generated at meiosis, including three gamete types of pure double reduction, partial double reduction, and no double reduction.

The statistical method proposed in this article describes a mapping framework for studying the genome structure and organization in complex autopolyploid species, providing a sophisticated model for linkage analysis in autopolyploids. It provides a necessary platform on which researchers can map quantitative trait loci (QTL) underlying economically and biologically important traits in autopolyploids. Although some preliminary studies have been reported for QTL mapping in autopolyploids, assuming pure bivalent pairings (Doerge and Craig 2000; Xie and Xu 2000), all of these should be viewed as premature until a comprehensive model is framed to take both bivalent and multivalent pairings into account.

We are grateful to Dr. S. Xu and Dr. M. Gallo-Meagher for stimulating discussions regarding this project. The authors thank the associate editor, referee Dr. R. Deborah Overath, and one anonymous referee for their constructive comments. This manuscript was approved as Journal Series R-08464 by the Florida Agricultural Experiment Station.

LITERATURE CITED

Bever, J. D., and F. Felber, 1992 The theoretical population genetics of autopolyploidy. Oxf. Surv. Evol. Biol. **8:** 185–217.

Brouwer, D. J., and T. C. Osborn, 1999 A molecular marker linkage map of tetraploid alfalfa (*Medicago sativa* L.). Theor. Appl. Genet. **99:** 1194–1200.

Darlington, C. D., 1929 Chromosome behaviour and structural hybridity in the Tradescantiae. J. Genet. **21:** 207–286.

da Silva, J., M. E. Sorrells, W. L. Burnquist and S. D. Tanksley, 1993 RFLP linkage map and genome analysis of *Saccharum spontaneum.* Genome **36:** 782–791.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via EM algorithm. J. Stat. Soc. Ser. B **39:** 1–38.

de Winton, D., and J. B. S. Haldane, 1931 Linkage in the tetraploid *Primula sinensis.* J. Genet. **24:** 121–144.

Doerge, R. W., and B. A. Craig, 2000 Model selection for quantitative trait locus analysis in polyploids. Proc. Natl. Acad. Sci. USA **97:** 7951–7956.

Fisher, R. A., 1947 The theory of linkage in polysomic inheritance. Philos. Trans. R. Soc. Ser. B **233:** 55–87.

Grant, V., 1981 *Plant Speciation*, Ed. 2. Columbia University Press, New York.

Grivet, L., A. D'Hont, D. Roques, P. Feldmann, C. Lanaud *et al.*, 1996 RFLP mapping in cultivated sugarcane (Saccharum spp): genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics **142:** 987–1000.

Hackett, C. A., J. E. Bradshaw, R. C. Meyer, J. W. McNicol, D. Milbourne *et al.*, 1998 Linkage analysis in tetraploid species: a simulation study. Genet. Res. **71:** 143–154.

Hilu, K. W., 1993 Polyploidy and the evolution of domesticated plants. Am. J. Bot. **80:** 1491–1499.

Jackson, R. C., and J. W. Jackson, 1996 Gene segregation in autotetraploids: prediction from meiotic configurations. Am. J. Bot. **83:** 673–678.

Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in human. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

Luo, Z. W., C. A. Hackett, J. E. Bradshaw, J. W. McNicol and D. Milbourne, 2000 Predicting parental genotypes and gene segregation for tetrasomic inheritance. Theor. Appl. Genet. **100:** 1067–1073.

Luo, Z. W., C. A. Hackett, J. E. Bradshaw, J. W. McNicol and D. Milbourne, 2001 Construction of a genetic linkage map in tetraploid species using molecular markers. Genetics **157:** 1369–1385.

Masterson, J., 1994 Stomatal size in fossil plants—evidence for polyploidy in majority of angiosperms. Science **264:** 421–424.

Mather, K., 1936 Segregation and linkage in autotetraploids. J. Genet. **32:** 287–314.

Meyer, R. C., D. Milbourne, C. A. Hackett, J. E. Bradshaw, J. W. McNichol and R. Waugh, 1998 Linkage analysis in tetraploid potato and association of markers with quantitative resistance to late blight (*Phytophthora infestans*). Mol. Gen. Genet. **259:** 150–160.

Otto, S. P., and J. Whitton, 2000 Polyploid incidence and evolution. Annu. Rev. Genet. **34:** 401–437.

Ripol, M. I., G. A. Churchill, J. A. G. da Silva and M. Sorrells, 1999 Statistical aspects of genetic mapping in autopolyploids. Gene **235:** 31–41.

Robert, P. C., and G. Casella, 2000 *Monte Carlo Statistical Methods* (Springer Texts in Statistics). Springer-Verlag, New York.

Self, S. G., and K. Y. Liang, 1987 Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard condition. J. Am. Stat. Assoc. **82:** 605–610.

Soltis, P. S., and D. E. Soltis, 2000 The role of genetic and genomic attributes in the success of polyploids. Proc. Natl. Acad. Sci. USA **97:** 7051–7057.

Stebbins, G. L., 1971 *Chromosomal Evolution in Higher Plants.* Addison-Wesley, Reading, MA.

Sved, J. A., 1964 The relationship between diploid and tetraploid recombination frequencies. Heredity **19:** 585–596.

Sybenga, A., 1994 Preferential pairing estimates from multivalent frequencies in tetraploids. Genome **37:** 1045–1055.

Wu, K. K., W. Burnquist, M. E. Sorrells, T. L. Tew, P. H. Moore *et al.*, 1992 The detection and estimation of linkage in polyploids using single-dose restriction fragments. Theor. Appl. Genet. **83:** L294–300.

Xie, C. G., and S. H. Xu, 2000 Mapping quantitative trait loci in tetraploid populations. Genet. Res. **76:** 105–115.

Yu, K. F., and K. P. Pauls, 1993 Segregation of random amplified polymorphic DNA markers and strategies for molecular mapping in tetraploid alfalfa. Genome **36:** 844–851.

Communicating editor: J. B. Walsh

APPENDIX

Among the 16 possible allele configurations, 4 have no recombination and 12 have one recombination. If we form gametes with two chromosomes by selecting, with replacement, from the 16 alleles twice, this yields $16 \times 16 = 256$ possibilities (16 with no recombination, 96 with one recombination, and 144 with two).

Recall that ϕ and ψ are the proportions of gamete types that have two recombination events under modes 7 and 8, respectively. Note that f_{7A} contains 12 out of 16 gametes with no recombination and f_{7B} contains 12 out of 144 gametes with two recombinations; thus the relative proportions should be $12(1-\theta)^2/16{:}12\theta^2/144 = 9(1-\theta)^2{:}\theta^2$. Similarly, f_{8A} contains 48 out of 96 gametes with one recombination and f_{8B} contains 48 out of 144 gametes with two recombinations; thus the relative proportions should be $48 \times 2\theta(1-\theta)/96{:}48 \times \theta^2/144 = 3(1-\theta){:}\theta$. Consequently, we may assume $\phi = \theta^2/(9(1-\theta)^2 + \theta^2)$ and $\psi = \theta/(3-2\theta)$.