DOI: 10.1534/genetics.105.044974

Modeling Population Genetic Data in Autotetraploid Species

Z. W. Luo,*,†,[1] Ze Zhang,* R. M. Zhang,† Madhav Pandey,‡ Oliver Gailing,‡ Hans H. Hattemer‡ and Reiner Finkeldey‡

**School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom,* †*Laboratory of Population and Quantitative Genetics, State Key Laboratory of Genetic Engineering, Fudan University, Shanghai 200433, China and* ‡*Institute of Forest Genetics and Forest Tree Breeding, Faculty of Forest Science and Forest Ecology, University of Göttingen, Büsgenweg 2, D-37077 Göttingen, Germany*

Manuscript received April 29, 2005
Accepted for publication September 13, 2005

ABSTRACT

Allozyme and PCR-based molecular markers have been widely used to investigate genetic diversity and population genetic structure in autotetraploid species. However, an empirical but inaccurate approach was often used to infer marker genotype from the pattern and intensity of gel bands. Obviously, this introduces serious errors in prediction of the marker genotypes and severely biases the data analysis. This article developed a theoretical model to characterize genetic segregation of alleles at genetic marker loci in autotetraploid populations and a novel likelihood-based method to estimate the model parameters. The model properly accounts for segregation complexities due to multiple alleles and double reduction at autotetrasomic loci in natural populations, and the method takes appropriate account of incomplete marker phenotype information with respect to genotype due to multiple-dosage allele segregation at marker loci in tetraploids. The theoretical analyses were validated by making use of a computer simulation study and their utility is demonstrated by analyzing microsatellite marker data collected from two populations of sycamore maple (*Acer pseudoplatanus* L.), an economically important autotetraploid tree species. Numerical analyses based on simulation data indicate that the model parameters can be adequately estimated and double reduction is detected with good power using reasonable sample size.

POLYPLOIDY has played an important role in the evolutionary diversification of up to 80% of angiosperm species (Grant 1971; Lewis 1980; Otto and Whitton 2000; Soltis and Soltis 2000). Two types of polyploids can be distinguished according to their genome origin. Allopolyploids are the product of an interspecific hybridization event and subsequent chromosome doubling, while autopolyploids originate from the whole-genome doubling, likely by fusion of unreduced conspecific gametes. Because bivalents are always formed between pairs of chromosomes with the same origin at meiosis, allopolyploids display disomic inheritance. In contrast, autopolyploids have more than two sets of homologous chromosomes, show multivalent chromosome pairing at meiosis, and display polysomic inheritance. The complexities in modeling polysomic inheritance lie in two major aspects: segregation of multiple dosage alleles at individual loci and the occurrence of double reduction, the phenomenon by which sister chromatids enter into the same gamete (Mather 1936). Another distinct feature in the population genetics of polyploids is the formation of partial heterozygotes. For example, when two alleles (A_1 and A_2) segregate at a locus in an autotetraploid population, there are three types of partial heterozygotes: $A_1A_1A_1A_2$, $A_1A_1A_2A_2$, and $A_1A_2A_2A_2$.

For their significance in evolutionary biology and agriculture, autopolyploids have attracted increasing research efforts at both theoretical and experimental scales (Ronfort *et al.* 1998; Luo *et al.* 2000, 2001, 2004; Mahy *et al.* 2000; Lopez-Pujol *et al.* 2004). Furthermore, rapid advances in the techniques of molecular biology and computer technology have made the genetical analysis of autopolyploids more tractable than ever before (De Winton and Haldane 1931; Fisher 1947). Allozyme and DNA-based microsatellite markers have been used to investigate the divergent heterozygosity between autotetraploids and their parental diploids (Mahy *et al.* 2000; Hardy and Vekemans 2001), to infer the genetic mode of autopolyploidy (Lopez-Pujol *et al.* 2004), and to assess population structure and gene flow in autotetraploid species (Ronfort *et al.* 1998). Thrall and Young (2000) developed a computer program, AUTOTET, for calculating allele frequencies of genetic markers in autotetraploid populations. However, it must be pointed out that to use the program, one has to empirically infer genotypes of the markers from inspecting the pattern and intensity of electrophoretic gel bands (for example, Lopez-Pujol *et al.* 2004). Obviously, diagnosing the intensity of gel bands could be unreliable or impossible for the marker data generated, for example, from DNA sequencers.

[1] *Corresponding author:* School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom.
E-mail: z.w.luo@bham.ac.uk

TABLE 1

The genotypic distribution at a single locus with *L* alleles in a random mating autotetraploid population

Zygotes	Genotypes	Probabilities	No. of terms
	Individuals carrying two DR gametes $[\times 9\alpha^2/(2+\alpha)^2]$		
Homozygotes	$A_iA_iA_iA_i$ $(1 \le i \le L)$	p_i^2	L
Heterozygotes	$A_iA_iA_jA_j$ $(1 \le i < j \le L)$	$2p_ip_j$	$L(L-1)/2$
	Individuals carrying one DR gamete $[\times 12\alpha(1-\alpha)/(2+\alpha)^2]$		
Homozygotes	$A_iA_iA_iA_i$ $(1 \le i \le L)$	p_i^3	L
Heterozygotes	$A_iA_iA_iA_j$ $(1 \le i < j \le L)$	$2p_i^2p_j$	$L(L-1)/2$
	$A_iA_jA_jA_j$ $(1 \le i < j \le L)$	$2p_ip_j^2$	$L(L-1)/2$
	$A_iA_iA_jA_j$ $(1 \le i < j \le L)$	$p_i^2p_j + p_ip_j^2$	$L(L-1)/2$
	$A_iA_iA_jA_k$ $(1 \le j < k \le L)$	$2p_ip_jp_k$	$L(L-1)(L-2)/6$
	$A_iA_jA_jA_k$ $(1 \le j < k \le L)$	$2p_ip_jp_k$	$L(L-1)(L-2)/6$
	$A_iA_jA_kA_k$ $(1 \le j < k \le L)$	$2p_ip_jp_k$	$L(L-1)(L-2)/6$
	Individuals carrying no DR gamete $[\times 4(1-\alpha)^2/(2+\alpha)^2]$		
Homozygotes	$A_iA_iA_iA_i$ $(1 \le i \le L)$	p_i^4	L
Heterozygotes	$A_iA_iA_iA_j$ $(1 \le i < j \le L)$	$4(p_i^3p_j + p_ip_j^3)$	$L(L-1)/2$
	$A_iA_iA_jA_j$ $(1 \le i < j \le L)$	$6p_i^2p_j^2$	$L(L-1)/2$
	$A_iA_iA_jA_k$ $(1 \le i < j,\ k \le L)$	$4p_i^2p_jp_k$	$L(L-1)(L-2)/6$
	$A_iA_jA_jA_k$ $(1 \le i < j,\ k \le L)$	$4p_ip_j^2p_k$	$L(L-1)(L-2)/6$
	$A_iA_jA_kA_k$ $(1 \le i < j,\ k \le L)$	$4p_ip_jp_k^2$	$L(L-1)(L-2)/6$
	$A_iA_jA_kA_l$ $(1 \le i < j,\ k < l \le L)$	$4p_ip_jp_kp_l$	$L(L-1)(L-2)(L-3)/4$

In this article, we develop a likelihood-based method for calculating allele frequencies of genetic markers in a random-mating autotetraploid population. The method accounts properly for the problem of missing information of marker phenotype in regard to the corresponding genotype and for the presence of double reduction. It can be used to analyze the genetic structure of autotetraploid populations by making use of allozyme and PCR-based molecular markers. The method is demonstrated by analyzing a data set consisting of five microsatellite markers scored on two populations of *Acer pseudoplatanus L.*, an autotetraploid tree species.

THEORY AND METHODS

Model and notation: We consider segregation of alleles at a locus in a random-mating autotetraploid population. Let L be the number of alleles, $A_1, A_2, \ldots, A_L$, segregating in the population and the frequencies of the alleles be $p_1, p_2, \ldots, p_L$, respectively. The coefficient of double reduction, which is the probability of sister chromatids ending up in the same gamete in meiosis (Mather 1936), at the marker locus is denoted by α. Geiringer (1949) found that the equilibrium distribution of zygotic genotypes is given by the expansion of

$$\frac{1}{(2+\alpha)^2}\left\{3\alpha\sum_{i=1}^{L}p_iA_i^2 + 2(1-\alpha)\left[\sum_{i=1}^{L}p_iA_i\right]^2\right\}^2. \quad (1)$$

If $\alpha = 0$, this is

$$(p_1 + p_2 + \cdots + p_L)^4 = \sum_{r_1+r_2+r_3+r_4=4}\frac{4!}{r_1!r_2!r_3!r_4!}p_i^{r_1}p_j^{r_2}p_k^{r_3}p_l^{r_4}$$
$$(1 \le i \ne j \ne k \ne l \le L). \quad (2)$$

After algebraic simplification, Equation 1 becomes

$$\begin{aligned}
&\frac{1}{(2+\alpha)^2}\left\{3\alpha\sum_{i=1}^{L}p_iA_i^2 + 2(1-\alpha)\left[\sum_{i=1}^{L}p_iA_i\right]^2\right\}^2\\
&= \frac{1}{(2+\alpha)^2}\left\{9\alpha^2\left[\sum_{i=1}^{L}p_i^2A_i^4 + 2\sum_{1\le i<j\le L}p_ip_jA_i^2A_j^2\right]\right.\\
&\quad + 12\alpha(1-\alpha)\left[\sum_{i=1}^{L}p_i^3A_i^4 + \sum_{1\le i\ne j\le L}p_ip_j^2A_i^2A_j^2\right.\\
&\quad\quad + 2\sum_{i=1}^{L}\sum_{\substack{j\ne i\\ 1\le j\le L}}p_i^2p_jA_i^3A_j\\
&\quad\quad \left. + 2\sum_{i=1}^{L}\sum_{\substack{j\ne i,k\ne i\\ 1\le j<k\le L}}p_ip_jp_kA_i^2A_jA_k\right]\\
&\quad + 4(1-\alpha)^2\left[\sum_{i=1}^{L}p_i^4A_i^4 + 6\sum_{1\le i<j\le L}p_i^2p_j^2A_i^2A_j^2\right.\\
&\quad\quad + 4\sum_{i=1}^{L}\sum_{\substack{j\ne i\\ 1\le j\le L}}p_i^3p_jA_i^3A_j + 12\sum_{1\le i<j\le L}p_i^2p_jp_kA_i^2A_jA_k\\
&\quad\quad \left.\left. + 24\sum_{i\ne j\ne k\ne l}p_ip_jp_kp_lA_iA_jA_kA_l\right]\right\}. \quad (3)
\end{aligned}$$

It can be seen that there are as many as $c(L,4) = L(L+1)(L+2)(L+3)/4!$ distinct genotypes at the marker locus (Luo and Ma 2004). To ease the following analysis, we classed the individual genotypes into three groups according to the number of double-reduction gametes they carried and illustrated the equilibrium genotypic distribution in Table 1.

TABLE 2

The phenotypic distribution at a single locus with *L* alleles in a random-mating autotetraploid population

Phenotypes (probability)	Form of the probability [$\times(2+\alpha)^{-2}$]	Observations	No. of terms
One band (f_i)	$9\alpha^2 p_i^2 + 12\alpha(1-\alpha)p_i^3 + 4(1-\alpha)^2 p_i^4$	n_i $(1 \le i \le L)$	L
Two bands (f_{ij})	$18\alpha^2 p_i p_j + 36\alpha(1-\alpha)\left[p_i^2 p_j + p_i p_j^2\right] + 4(1-\alpha)^2\left[4p_i^3 p_j + 6p_i^2 p_j^2 + 4p_i p_j^3\right]$	n_{ij} $(1 \le i < j \le L)$	$L(L-1)/2$
Three bands (f_{ijk})	$72\alpha(1-\alpha)p_i p_j p_k + 48(1-\alpha)^2 \times \left[p_i^2 p_j p_k + p_i p_j^2 p_k + p_i p_j p_k^2\right]$	n_{ijk} $(1 \le i < j < k \le L)$	$L(L-1)(L-2)/6$
Four bands (f_{ijkl})	$96(1-\alpha)^2 p_i p_j p_k p_l$	n_{ijkl} $(1 \le i < j < k < l \le L)$	$L(L-1)(L-2)(L-3)/24$

Analysis and statistical inference: The genotypic distribution involves L independent unknown parameters: $\alpha, p_1, p_2, \ldots, p_{L-1}$ since $\sum_i^L p_i = 1$. When the individual genotypes are distinguishable directly from the phenotype data, estimation of the double-reduction and allelic frequency parameters becomes trivial. However, there is no one-to-one relationship between genotype and phenotype at any genetic markers in the autotetraploids (Luo *et al.* 2000). To estimate the model parameters, one has to work with phenotype data. For simplicity but without loss of generality, we denote the phenotype of an autotetraploid individual at a marker locus by the number of distinct gel bands scored at the marker locus. For a given dominance mode of the markers, it is easy to convert the marker genotype distribution into the corresponding phenotype distribution. For example, the genotype distribution tabulated in Table 1 can be transformed into the phenotype distribution that is summarized in Table 2 when the marker alleles are assumed to display codominant inheritance. It can be seen from Table 2 that a phenotype provides full information of the corresponding genotype only if the phenotype is represented as four distinct bands; *i.e.*, the presence of four different alleles and, in most cases, an individual genotype cannot be inferred directly from the corresponding phenotype. A general form for converting the genotypic distribution (Table 1) into the phenotypic distribution (Table 2) is $f_x = \sum_{h \in x} g_h$, where the summation is over all genotypes that correspond to the same phenotype, $x = i$ (one band), ij (two bands), ijk (three bands), or $ijkl$ (four bands). g_h is the probability of genotype h that is compatible with phenotype x. For example, there are three genotypes ($A_iA_iA_iA_j$, $A_iA_iA_jA_j$, and $A_iA_jA_jA_j$) that are compatible with phenotype ij; thus f_{ij}, the probability of this phenotype, equals $\{18\alpha^2 p_i p_j + 36\alpha(1-\alpha)[p_i^2 p_j + p_i p_j^2] + 4(1-\alpha)^2[4p_i^3 p_j + 6p_i^2 p_j^2 + 4p_i p_j^3]\}/(2+\alpha)^2$ as given in Table 2. Let n be the number of individuals randomly sampled from the population under question and m be the number of different phenotypes observed in the sample. The sample can be divided into phenotype groups for which there are n_i, n_{ij}, n_{ijk}, and n_{ijkl} of individuals with one, two, three, and four bands, accordingly. The logarithm of the likelihood function of the observed phenotype data given the model parameters is given by

$$\begin{aligned}\log[\ell(\alpha, p_1, \ldots, p_{L-1} | n_i, n_{ij}, n_{ijk}, n_{ijkl})] \propto & \sum_i n_i \log(f_i) \\ & + \sum_{ij} n_{ij} \log(f_{ij}) \\ & + \sum_{ijk} n_{ijk} \log(f_{ijk}) \\ & + \sum_{ijkl} n_{ijkl} \log(f_{ijkl}) \\ & - 2n \log(2+\alpha).\end{aligned} \tag{4}$$

Differentiating Equation 4 with respect to α gives

$$\begin{aligned}\frac{\partial}{\partial\alpha}\log[\ell(\alpha, p_1, \ldots, p_{L-1} | n_i, n_{ij}, n_{ijk}, n_{ijkl})] &= \frac{1}{\alpha(1-\alpha)}\Big[\sum_i n_i(2\xi_{i1} + \xi_{i2}) + \sum_{ij} n_{ij}(2\xi_{ij1} + \xi_{ij2}) \\ &\quad + \sum_{ijk} n_{ijk}\xi_{ijk1} - 2n\alpha\Big] \\ &\quad - \frac{2n}{2+\alpha},\end{aligned} \tag{5}$$

where

$$\begin{aligned}\xi_{i1} &= \frac{9\alpha^2 p_i^2}{f_i} \\ \xi_{i2} &= \frac{12\alpha(1-\alpha)p_i^3}{f_i} \\ \xi_{ij1} &= \frac{18\alpha^2 p_i p_j}{f_{ij}} \\ \xi_{ij2} &= \frac{36\alpha(1-\alpha)[2p_i^2 p_j + p_i p_j^2]}{f_{ij}} \\ \xi_{ijk1} &= \frac{72\alpha(1-\alpha)p_i p_j p_k}{f_{ijk}}\end{aligned} \tag{6}$$

in which f_i, f_{ij}, and f_{ijk} are given in Table 2. The maximum-likelihood estimate (MLE) of the unknown parameter, α, can be solved from setting the derivative equation to zero as

$$\hat{\alpha} = \frac{2c}{6n - c} \tag{7}$$

in which $c = \sum_i n_i(2\xi_{i1} + \xi_{i2}) + \sum_{ij} n_{ij}(2\xi_{ij1} + \xi_{ij2}) + \sum_{ijk} n_{ijk}\xi_{ijk1}$. Differentiating Equation 4 with respect to $p_i (i = 1, 2, \ldots, L-1)$ gives

$$\frac{\partial}{\partial p_i}\log[\ell(\alpha, p_1, \ldots, p_{L-1}|n_i, n_{ij}, n_{ijk}, n_{ijkl})] = \frac{1}{p_i p_L}[c_{iL}p_L - c_i p_i], \tag{8}$$

where

$$\begin{aligned} c_i = {} & n_L(2 + \eta_{L2} + 2\eta_{L3}) + n_{iL}(1 + \eta_{iL3} + \eta_{iL5} + 2\eta_{iL6}) \\ & + \sum_{j\neq i} n_{jL}(1 + \eta_{jL3} + \eta_{jL5} + 2\eta_{jL6}) \\ & + \sum_{i<j} n_{ijL}(1 + \eta_{ijL4}) + \sum_{j<i} n_{jiL}(1 + \eta_{jiL4}) + \sum_{j<k} n_{jkL}(1 + \eta_{jkL4}) \\ & + \sum_{j<k}(n_{ijkL} + n_{jikL} + n_{jkiL}) \\ & + \sum_{j<k<l} n_{jklL} \end{aligned} \tag{9}$$

and

$$\begin{aligned} c_{iL} = {} & n_i(2 + \eta_{i2} + 2\eta_{i3}) + \sum_{1\leq i<j} n_{ij}(1 + \eta_{ij2} + 2\eta_{ij4} + \eta_{ij5}) \\ & + \sum_{1\leq j<i} n_{ji}(1 + \eta_{ij3} + \eta_{ij5} + 2\eta_{ij6}) + n_{iL}(1 + \eta_{iL2} + 2\eta_{iL4} + \eta_{iL5}) \\ & + \sum_{i<j<k} n_{ijk}(1 + \eta_{ijk2}) + \sum_{j<i<k} n_{jik}(1 + \eta_{jik3}) + \sum_{j<k<i} n_{jki}(1 + \eta_{jki4}) \\ & + \sum_{i<j} n_{ijL}(1 + \eta_{ijL2}) + \sum_{j<i} n_{jiL}(1 + \eta_{jiL3}) + \sum_{i<j<k<l} n_{ijkl} \\ & + \sum_{j<i<k<l} n_{jikl} + \sum_{j<k<i<l} n_{jkil} + \sum_{j<k<l<i} n_{jkli} \\ & + \sum_{j<k}(n_{ijkL} + n_{jikL} + n_{jkiL}) \end{aligned} \tag{10}$$

in which

$$\begin{aligned} \eta_{i2} &= \frac{12\alpha(1-\alpha)p_i^3}{f_i}, & \eta_{i3} &= \frac{4(1-\alpha)^2 p_i^4}{f_i} \\ \eta_{ij2} &= \frac{36\alpha(1-\alpha)p_i^2 p_j}{f_{ij}}, & \eta_{ij3} &= \frac{36\alpha(1-\alpha)p_i p_j^2}{f_{ij}} \\ \eta_{ij4} &= \frac{16(1-\alpha)^2 p_i^3 p_j}{f_{ij}}, & \eta_{ij5} &= \frac{24(1-\alpha)^2 p_i^2 p_j^2}{f_{ij}} \\ \eta_{ij6} &= \frac{16(1-\alpha)^2 p_i p_j^3}{f_{ij}}, & \eta_{ijk2} &= \frac{48(1-\alpha)^2 p_i^2 p_j p_k}{f_{ijk}} \\ \eta_{ijk3} &= \frac{48(1-\alpha)^2 p_i p_j^2 p_k}{f_{ijk}}, & \eta_{ijk4} &= \frac{48(1-\alpha)^2 p_i p_j p_k^2}{f_{ijk}} \end{aligned} \tag{11}$$

Setting the derivative Equations 8 to equal zero, we obtain $L-1$ linear equations in the form of

$$\mathbf{AP} = \mathbf{B}, \tag{12}$$

where $\mathbf{A} = (a_{ij})_{L-1\times L-1}$ is a square matrix with $a_{ii} = c_{iL} + c_i$ and $a_{ij} = c_{iL}$, $\mathbf{B} = (c_{1L}, c_{2L}, \ldots, c_{L-1L})^{\mathrm{T}}$, and $\mathbf{P} = (p_1, p_2, \ldots, p_{L-1})^{\mathrm{T}}$. Solving Equation 12 yields the MLEs of allelic frequencies $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{L-1}$. In fact, Equations 6 and 7 and Equations 11 and 12 represent an EM algorithm for calculating the maximum-likelihood estimates of the model parameters. The algorithm involves iterating the two steps: the expectation step that calculates the posterior probabilities according to Equations 6 and 11 and then the maximization step that calculates the maximum-likelihood estimates of the model parameters from Equations 7 and 12 with ξ's and η's being calculated from the previous expectation step. These two steps are iterated until the sequence of the likelihood function (4) converges. In contrast, the MLEs of marker allele frequencies can be simplified as $\hat{p}_i = (2n_i + \sum_{j\neq i} n_{ij})/2n$ in diploid populations.

It is feasible to create a likelihood-based statistical test for the presence of double reduction with the MLEs, $\hat{\alpha}, \hat{p}_1, \ldots, \hat{p}_{L-1}$. In fact, the test statistic given by

$$\chi^2_{\mathrm{d.f.}=1} = 2\log[\ell(\hat{\alpha}, \hat{p}_1, \ldots, \hat{p}_{L-1}|n_i, n_{ij}, n_{ijk}, n_{ijkl})/\ell(0, \hat{p}_1, \ldots, \hat{p}_{L-1}|n_i, n_{ij}, n_{ijk}, n_{ijkl})] \tag{13}$$

has an asymptotic chi-square distribution with 1 d.f. and can be used to test the significance of double reduction at the locus under question.

Test for Hardy–Weinberg hypothesis: The above analysis is based on the assumption that segregation of the marker alleles follows the Hardy–Weinberg equilibrium; *i.e.*, frequencies of marker genotypes are determined by the coefficient of double reduction and the relevant allele frequencies. With the MLEs of the model parameters, the expected genotype (or phenotype) frequencies can be calculated and thus a chi-square statistic with a form of

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \tag{14}$$

can be constructed to test for significance of the hypothesis. It should be pointed out that the number of segregating alleles in the population might be so large for the sample size that the numbers of individuals for most genotype classes are too small for the chi-square test to be reliable. To avoid this problem, we suggest that the expected number of the phenotype classes that is >1 be used to calculate the chi square.

Estimation of genetic heterozygosity: With the MLEs of the allelic frequencies, one can estimate genetic heterozygosity at marker loci. The genetic heterozygosity, which accounts for the distinct allelic constitutions of tetraploid genotypes, was defined as the probability of any two alleles being not identical in a given genotype (Bever and Felber 1992; Thrall and Young 2000). In the present context and notations, the observed genetic heterozygosity can be calculated from

$$\begin{aligned} H_{\mathrm{O}} = \frac{1}{6n}\Big\{ & \sum 6n_{ijkl} + \sum 5n_{ijk} + \sum n_{ij}[3(\eta_{ij2} + \eta_{ij3} + \eta_{ij4} + \eta_{ij6}) \\ & + 4(\eta_{ij1} + \eta_{ij5})]\Big\}, \end{aligned} \tag{15}$$

whose expectation is given by

TABLE 3

Means and corresponding standard deviations (in parentheses) of the maximum-likelihood estimates of simulated coefficient of double reduction, $\alpha = 0.05$ or 0.1, and allelic frequencies p_i ($i = 1, 2, \ldots, 5$) over 100 repeated simulations of varying sample sizes, n

n	$p_1 = 0.2$	$p_2 = 0.2$	$p_3 = 0.2$	$p_4 = 0.2$	$p_5 = 0.2$	$\hat{\alpha}$	$\hat{\beta}$ (%)
50	0.2014 (0.034)	0.1971 (0.033)	0.1988 (0.032)	0.2035 (0.033)	0.1993 (0.031)	0.0586 (0.053)	14
100	0.2012 (0.025)	0.2005 (0.023)	0.2006 (0.024)	0.1968 (0.023)	0.2009 (0.025)	0.0469 (0.035)	12
200	0.2007 (0.017)	0.1981 (0.016)	0.2008 (0.015)	0.1983 (0.017)	0.2021 (0.015)	0.0466 (0.30)	29
50	0.1980 (0.031)	0.1948 (0.031)	0.1972 (0.032)	0.2051 (0.034)	0.2049 (0.032)	0.0954 (0.062)	31
100	0.1980 (0.024)	0.2020 (0.024)	0.1971 (0.025)	0.1989 (0.020)	0.2042 (0.024)	0.0904 (0.045)	54
200	0.1998 (0.017)	0.1990 (0.017)	0.2005 (0.017)	0.2016 (0.017)	0.1991 (0.017)	0.0943 (0.032)	84
	$p_1 = 0.4$	$p_2 = 0.3$	$p_3 = 0.19$	$p_4 = 0.1$	$p_5 = 0.01$		
50	0.4011 (0.044)	0.3008 (0.039)	0.1929 (0.030)	0.0955 (0.022)	0.0097 (0.006)	0.0523 (0.051)	4
100	0.3999 (0.030)	0.2996 (0.029)	0.1918 (0.020)	0.0992 (0.016)	0.0095 (0.005)	0.0581 (0.040)	10
200	0.4009 (0.023)	0.3007 (0.020)	0.1894 (0.014)	0.0995 (0.013)	0.0095 (0.003)	0.0514 (0.034)	23
50	0.4043 (0.044)	0.2956 (0.046)	0.1872 (0.030)	0.1027 (0.025)	0.0101 (0.007)	0.0979 (0.069)	22
100	0.4026 (0.030)	0.2973 (0.025)	0.1888 (0.022)	0.1013 (0.016)	0.0100 (0.005)	0.0934 (0.054)	44
200	0.4015 (0.023)	0.3022 (0.019)	0.1872 (0.016)	0.0983 (0.011)	0.0109 (0.004)	0.0912 (0.042)	65

$\hat{\beta}$ is the empirical power of the statistical test for double reduction.

$$H_E = \frac{1}{(2+\alpha)^2}\Bigg\{\sum_{1\le i<j\le L}(12\alpha^2 p_i p_j + 16(1-\alpha)^2 p_i^2 p_j^2) + \sum_{1\le i<j\le L}[18\alpha(1-\alpha)(p_i^2 p_j + p_i p_j^2) + 8(1-\alpha)^2(p_i^3 p_j + p_i p_j^3)] + \sum_{1\le i<j<k\le L}[60\alpha(1-\alpha)p_i p_j p_k + 40(1-\alpha)^2(p_i^2 p_j p_k + p_i p_j^2 p_k + p_i p_j p_k^2)] \times \sum_{1\le i<j<k<l\le L} 96(1-\alpha)^2 p_i p_j p_k p_l\Bigg\}. \quad (16)$$

When marker data are collected from k subpopulations of a tetraploid species, one can calculate marker heterozygosity for each of these populations according to the above analysis. Let H_i be the heterozygosity in the subpopulation I, and $H_I = \sum_{i=1}^{k} H_i/k$. We define forms of Wright's F-statistics for autotetraploids by assimilating that in diploids (Hartl and Clark 1997) as

$$F_{IS} = \frac{\bar{H}_S - H_I}{\bar{H}_S} \quad (17)$$

and

$$F_{ST} = \frac{H_T - \bar{H}_S}{H_T}, \quad (18)$$

where $H_S = \sum_{s=1}^{k} H_E/k$ and H_T is the expected heterozygosity of an equivalent total population of all the subpopulations under the Hardy–Weinberg equilibrium.

NUMERICAL ANALYSES

Simulation study: To validate the method presented above and to investigate its properties, we set up a simulation study to mimic segregation of alleles at a single marker locus in autotetraploid populations at equilibrium. The simulation model allows us to vary the number and frequencies of marker alleles and to simulate different values of the coefficient of double reduction. For any given simulated parameters, individual genotypes were randomly sampled from the simulated population for which the genotypic distribution is defined in Table 1. The sampled genotypes were converted into phenotypes according to Luo *et al* (2000). The simulation study considered the segregation of five marker alleles with either an equal frequency or different frequencies in populations of different sample sizes and two values of the coefficient of double reduction, $\alpha = 0.05$ and 0.1, at the marker locus.

Tabulated in Table 3 are the means and standard deviations of the MLEs of the coefficient of double reduction and allelic frequencies over 100 repeated simulations, together with the simulation parameters. Also, empirical power, $\hat{\beta}$, is listed for detecting significance of double reduction, which was calculated as the proportion of significant tests of double reduction in the repeated simulations. It can be seen from Table 3 that the parameters are estimated adequately in all cases even with a sample size as small as $n = 50$. Double reduction can be detected with fairly good power when the sample size is not <100 and when $\alpha = 0.1$. For a given sample size, double reduction is tested with larger power when marker alleles are evenly distributed as compared to noneven distribution of marker alleles. As expected, the standard deviation of the estimates decreases and the power of the double reduction test increases with increasing sample size.

Analysis of microsatellite data from two sycamore maple populations: Sycamore maple (*A. pseudoplatanus* L.) is a forest tree species native to central Europe and

TABLE 4

Summary and analysis of microsatellite marker data collected from two Sycamore maple (*Acer pseudoplatanus*) populations, Södderich and Weisswassertal

Södderich ($n = 133$)					Weisswassertal ($n = 80$)				
Phenotype	Obs.	Exp.	Allele and $\hat{p}_i$		Phenotype	Obs.	Exp.	Allele and $\hat{p}_i$	
1 0 1 0 0 0 0	81	67.14	1	0.4390	1 1 0 0 0	49	42.69	1	0.4418
1 0 1 1 0 0 0	27	19.30	2	0.0038	1 1 1 0 0	14	9.20	3	0.4418
1 0 1 0 0 1 0	4	6.72	3	0.4325	1 1 0 1 0	7	6.63	4	0.0525
1 0 1 0 1 0 0	4	5.64	4	0.0678	1 1 0 0 1	5	4.33	5	0.0384
1 1 1 0 0 0 0	2	1.00	5	0.0209	1 1 1 1 0	2	0.76	6	0.0254
1 0 1 0 0 0 1	1	3.02	6	0.0247	1 1 0 1 1	3	0.37		
1 0 0 1 0 1 0	1	0.62	7	0.0113					
1 0 1 0 1 1 0	6	0.31							
1 0 1 1 0 0 1	4	0.46							
1 0 1 1 1 0 0	1	0.86							
1 0 1 1 0 1 0	1	1.02	$\hat{\alpha} = 0.0000$					$\hat{\alpha} = 0.0001$	
1 0 1 0 0 1 1	1	0.17	$\chi^2_{\text{d.f.}=6} = 9.86$					$\chi^2_{\text{d.f.}=3} = 3.56$	

$\hat{\alpha}$ and $\hat{p}_i$ are the maximum-likelihood estimates of the coefficient of double reduction and frequency of the ith marker allele. The chi-square value, $\chi^2_{\text{d.f.}}$, was calculated by sorting together those phenotype classes whose expected counts were <1. Obs., observed; Exp., expected.

western Asia, growing especially in mountainous regions. The species is an autotetraploid with a chromosome number of $2n = 4x = 52$ (Darlington and Wylie 1955). Because of its economic and ecological importance, numerous plantations were established throughout Europe during the last century. However, the genetic structure and its mating system in natural populations or in plantations remain unclear.

To explore the genetic variation pattern, gene flow, and mating system of the species, Pandey *et al.* (2004) have started developing microsatellite markers in this species. For a proof-of-principle demonstration purpose, we detailed here analysis of data of a marker, MAP-9, which was scored for 133 and 80 individuals, respectively, from two populations, Södderich and Weisswassertal, near Göttingen in Germany. The marker data, which were collected as peak value reads from the ABI PRISM 3100 genetic analyzer (Applied Biosystems, Foster City, CA/Hitachi, San Jose, CA), were converted into gel-band-like phenotypes.

Table 4 summarizes the phenotypic distribution and analysis of the marker data in the two sycamore maple populations. It shows that the 133 and 80 individuals sampled from the two populations can be grouped into 12 and 6 different phenotypic groups, respectively. From the phenotype, seven distinct marker alleles are observed in the Södderich population sample but only five of these are present in the Weisswassertal population sample. From the data sets, the maximum-likelihood estimates are calculated for frequencies of the marker alleles and the coefficient of double reduction by making use of the method developed in this study. It shows that marker alleles 1 and 3 are detected in the two populations as the most frequently occurring alleles. There is no evidence of double reduction at this marker locus. The goodness-of-fit test demonstrates that segregation of the marker alleles in both the populations agrees well with the Hardy–Weinberg equilibrium ($\chi^2_{\text{d.f.}=6} = 9.86, P > 0.13$; $\chi^2_{\text{d.f.}=3} = 3.56, P > 0.3$). To make the chi-square-based fitness tests more stable (Kendall and Stuart 1961, p. 440), the analysis has excluded those phenotype classes with expected sample observations <1.0. In addition, we calculated genetic heterozygosity (H_O) at the marker locus using Equation 15 and observed that $H_O = 0.6901$ and 0.6833 for the Södderich and the Weisswassertal populations, respectively. On the basis of the observed genetic heterozygosity estimates and their corresponding expected values that can be calculated from Equation 16, the F_{ST} is estimated to be 0.001, suggesting trivial genetic differentiation between the two populations.

DISCUSSION

There has been increasing interest in investigating genetic diversity and population genetic structure of autotetraploid species for understanding evolutionary impacts of polyploidy and the conservation biology of the species (Brown and Young 2000; Mahy *et al.* 2000; Hardy and Vekemans 2001; Lopez-Pujol *et al.* 2004). This has been greatly promoted by the increasing availability of allozyme and PCR-based DNA molecular marker data in these populations. However, analysis of the fast growing data sets of DNA polymorphic markers raises challenges to the development of appropriate statistical methods and algorithms. The data sets have been analyzed by empirically inferring individual

genotype from the intensity of gel bands (Soltis and Soltis 1989; Wendel and Weeden 1989; Lopez-Pujol *et al.* 2004). It is well known that this will be either inaccurate or unfeasible for most PCR-based DNA markers. Thus, development of theory and methods for modeling and analyzing these data sets meets the urgent needs of the research area.

This article developed a theoretical model for characterizing genetic segregation of alleles at genetic marker loci, an autotetraploid population locus and a novel maximum-likelihood-based method to estimate the parameters defining the population genetic model. The model properly accounts for segregation complexities due to multiple alleles and double reduction at tetrasomic loci in natural populations, and the statistical method takes appropriate account of incomplete marker phenotype information with respect to the genotype due to multiple dosage allele segregation at marker loci in tetraploids. The theoretical analysis and methods developed in this study are validated by making use of a computer simulation study and their utility is demonstrated by analyzing microsatellite marker data collected from two populations of sycamore maple (*A. pseudoplatanus* L.), an economically important autotetraploid tree species. Numerical analyses based on simulation data indicate that the model parameters can be adequately estimated and double reduction is detected with good power with a sample size of ~100. Analysis of the sycamore maple data illustrates the data format required by the statistical method and demonstrates how the method can be used to predict various population genetic parameters by using DNA molecular data in autotetraploid species. A practical problem in real data analysis might be that the number of segregating alleles is too large in relation to reasonable sample sizes. Thus, the sample size required for accurate parameter estimation and efficient testing of double reduction may depend largely on the number of segregating alleles; the sample size considered here may be suitable only when there are several marker alleles segregating in the population.

Although the theory and methods developed in this study were demonstrated for codominant markers, it is feasible for their extension to dominant markers such as RAPDs, AFLPs, etc. In fact, there are usually two segregating alleles at these marker loci, *i.e.*, a dominant and a recessive allele; this greatly reduces the number of possible genotypes and the number of possible phenotypes at the locus and thus simplifies the model. In addition, the recessive allele will be phenotypically present only when it is present in four copies.

Double reduction is one of the most distinguished features of tetrasomic inheritance and also one of the major difficulties in modeling tetrasomic inheritance when compared to disomic inheritance. Butruille and Boiteux (2000) exploited the impact of double reduction on the evolution of deleterious mutants in autotetraploid genomes and found that low frequencies of double reduction are enough to reduce equilibrium frequencies, which are maintained at a selection and mutation balance, by severalfold. This suggests the important effect of double reduction on the evolution of the species. However, there are no proposals, either in their study or in the literature, to address how to test for the significance of double reduction in natural populations of the species. This study provides such a test and thus fills a gap in the evolutionary study of autotetraploid species.

The algorithm developed in this study was programmed in Fortran-90 computer language. The program is available upon request from the corresponding author.

We thank Lindsey Leach and two anonymous reviewers for reading and commenting on the manuscript. This study was supported by research grants from the Biotechnology and Biological Sciences Research Council and the Natural Environment Research Council. Z.W.L. and R.M.Z. were also supported by the National Natural Science Foundation (30430380) and Shanghai Science & Technology Committee. M. Pandey was supported by the Deutsche Forschungsgemeinschaft (Ha 501/32-1).

LITERATURE CITED

Bever, J. D., and F. Felber, 1992 The theoretical population genetics of autopolyploidy, pp. 185–217 in *Oxford Surveys in Evolutionary Biology*, edited by J. Antonovics and D. Futuyma. Oxford University Press, Oxford.

Brown, A. H. D., and A. G. Young, 2000 Genetic diversity in tetraploid populations of the endangered daisy *Rutidosis leptorrhynchoides* and implications for its conservation. Heredity **85:** 122–129.

Butruille, D. V., and L. S. Boiteux, 2000 Selection-mutation balance in polysomic tetraploids: impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. Proc. Natl. Acad. Sci. USA **97:** 6608–6613.

Darlington, C. D., and A. P. Wylie, 1955 *Chromosome Atlas of Flowering Plants.* George Allen & Unwin, London.

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

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

Geiringer, H., 1949 Chromatid segregation in tetraploids and hexaploids. Genetics **34:** 665–684.

Grant, V., 1971 *Plant Speciation.* Columbia University Press, New York/London.

Hardy, O. J., and X. Vekemans, 2001 Patterns of allozyme variation in diploid and tetraploid *Centaurea jacea* at different spatial scales. Evolution **55:** 943–954.

Hartl, D. H., and A. G. Clark, 1997 *Principles of Population Genetics*, Ed. 3. Sinauer Associates, Sunderland, MA.

Kendall, M. G., and A. Stuart, 1961 *The Advanced Theory of Statistics: Inference and Relationship*, Vol. 2. Charles Griffin, London.

Lewis, W. H., 1980 *Polyploidy: Biological Relevance.* Plenum Press, New York.

Lopez-Pujol, J., M. Bosch, J. Simon and C. Blanche, 2004 Allozyme diversity in the tetraploid endemic *Thymus loscossi* (Lamiaceae). Ann. Bot. **93:** 323–332.

Luo, Z. W., and L. Ma, 2004 An improved formulation of marker heterozygosity in recurrent selection and backcross schemes. Genet. Res. **83:** 49–53.

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.

Luo, Z. W., R. M. Zhang and M. J. Kearsey, 2004 Theoretical basis for genetic linkage analysis in autotetraploid species. Proc. Natl. Acad. Sci. USA **101:** 7040–7045.

Mahy, G., L. P. Bruederle, B. Connors, M. Van Hofwegen and N. Vorsa, 2000 Allozyme evidence for genetic autopolyploidy and high genetic diversity in tetraploid cranberry, *Vaccinium oxycoccos* (Ericaceae). Am. J. Bot. **87:** 1882–1889.

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

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

Pandey, M., O. Gailing, D. Fischer, H. H. Hattemer and R. Finkeldey, 2004 Characterization of microsatellite markers in sycamore (*Acer pseudoplatanus* L.). Mol. Ecol. Notes **4:** 253–255.

Ronfort, J., E. Jenczewski, T. Bataillon and F. Rousset, 1998 Analysis of population structure in autotetraploid species. Genetics **150:** 921–930.

Soltis, D. E., and P. S. Soltis, 1989 Genetic consequences of autopolyploidy in *Tolmiea* (Saxifragaceae). Evolution **43:** 586–594.

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.

Thrall, P. H., and A. Young, 2000 AUTOTET: a program for the analysis of autotetraploid genotypic data. J. Hered. **91:** 348–349.

Wendel, F., and N. F. Weeden, 1989 Visualization and interpretation of plant isozymes, pp. 5–45 in *Isozymes in Plant Biology*, edited by D. E. Soltis and P. S. Soltis. Dioscorides Press, Portland, OR.

Communicating editor: Y.-X. Fu