Construction of a Genetic Linkage Map in Tetraploid Species Using Molecular Markers

Z. W. Luo,*,† C. A. Hackett,‡ J. E. Bradshaw,§ J. W. McNicol‡ and D. Milbourne§

**School of Biosciences, The University of Birmingham, Birmingham B15 2TT, England, †Laboratory of Population and Quantitative Genetics, Institute of Genetics, Fudan University, Shanghai 200433, China and ‡Biomathematics and Statistics Scotland, §Scottish Crop Research Institute, Invergowrie, Dundee, DD2 5DA, Scotland*

Manuscript received May 1, 2000
Accepted for publication November 30, 2000

ABSTRACT

This article presents methodology for the construction of a linkage map in an autotetraploid species, using either codominant or dominant molecular markers scored on two parents and their full-sib progeny. The steps of the analysis are as follows: identification of parental genotypes from the parental and offspring phenotypes; testing for independent segregation of markers; partition of markers into linkage groups using cluster analysis; maximum-likelihood estimation of the phase, recombination frequency, and LOD score for all pairs of markers in the same linkage group using the EM algorithm; ordering the markers and estimating distances between them; and reconstructing their linkage phases. The information from different marker configurations about the recombination frequency is examined and found to vary considerably, depending on the number of different alleles, the number of alleles shared by the parents, and the phase of the markers. The methods are applied to a simulated data set and to a small set of SSR and AFLP markers scored in a full-sib population of tetraploid potato.

GENETIC linkage maps are now available for man and for a large number of diploid plant and animal species. In contrast, mapping studies in polyploid species are much less advanced, partly due to the complexities in analysis of polysomic inheritance as demonstrated in, for example, Mather (1936), De Winton and Haldane (1931), Fisher (1947), and Bailey (1961). The development of DNA molecular markers [restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), randomly amplified polymorphic DNAs (RAPDs), simple sequence repeats (SSRs), and single nucleotide polymorphisms (SNPs), etc.] and advances in computer technology have made both theoretical and experimental studies of polysomic inheritance much more feasible than ever before. Some of these markers have recently been used as a fundamental tool to construct genetic linkage maps in polyploid species that display polysomic inheritance (Al-Janabi *et al.* 1993; Da Silva *et al.* 1993; Yu and Pauls 1993; Hackett *et al.* 1998; Brouwer and Osborn 1999), to search for quantitative trait loci (QTL) affecting disease resistance in tetraploid potato (Bradshaw *et al.* 1998; Meyer *et al.* 1998), and to investigate population structure in autotetraploid species (Ronfort *et al.* 1998).

Due to a lack of well-established theory for mapping genetic markers in polyploid species, much research has been based on strategies by which the complexities involved in modeling polysomic inheritance can be avoided. These involve either the use of single-dose (simplex) dominant markers (*e.g.*, AFLPs and RAPDs) that segregate in a simple 1:1 ratio in segregating populations or use of the corresponding diploid relative as an approximation to the polyploid case (Bonierbale *et al.* 1988; Gebhardt *et al.* 1989). More recently, Hackett *et al.* (1998) presented a theoretical and simulation study on linkage analysis of dominant markers of different dosages in a full-sib population of an autotetraploid species, and this approach was used by Meyer *et al.* (1998) to develop a linkage map in tetraploid potato.

The use of codominant markers, particularly those with a high degree of polymorphism such as SSRs, is known to improve the efficiency and accuracy of linkage analysis in diploid species (Terwilliger *et al.* 1992; Jiang and Zeng 1997). In polyploid species, the relationship between the parental genotype and the phenotype as shown by the gel band pattern is less clear-cut, due to the possibilities of different dosages of alleles, and this provides extra complexity as explained in Luo *et al.* (2000). The aim of the present study is to develop methodology for constructing linkage maps of codominant or dominant genetic markers in autotetraploid species under chromosomal segregation, *i.e.*, the random pairing of four homologous chromosomes to give two bivalents. The complications arising from quadrivalent or trivalent plus univalent formation are not considered in this article. A series of problems involved in tetrasomic linkage analysis are addressed. Statistical properties of the methods are investigated by theoretical

Corresponding author: Z. W. Luo, School of Biosciences, The University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom. E-mail: zwluo@bham.ac.uk

analysis or simulation study, and some experimental data from a tetraploid potato study are used to illustrate the use of the theory and methods in analyzing breeding experiments.

THEORY OF LINKAGE MAP CONSTRUCTION

Model and notation: The theoretical analysis considers a full-sib family derived from crossing two autotetraploid parental lines. Let M_i ($i = 1 \ldots m$) be m marker loci (with dominant or codominant inheritance). Let G_1 and G_2 be the genotypes at the marker loci for two parental individuals, respectively. G_i ($i = 1, 2$) can be expressed as a $m \times 4$ matrix. Because two tetraploid individuals have at most eight distinct alleles, we represent each element of G_i as a letter A–H or O, where O represents the null allele due to mutation within primer sequences (see, for example, Callen *et al.* 1993). It is important to note that allele A at marker locus 1 is different from allele A at marker locus 2.

When we are considering linked loci, it is often necessary to specify how the alleles at different loci are grouped into homologous chromosomes, *i.e.*, the linkage phases of the alleles. Alleles linked on the same homologous chromosome will appear in the same column of the matrix G_i. For a two-locus genotype with four different alleles at each locus, one possible genotype is

$$\begin{pmatrix} ABCD \\ ABCD \end{pmatrix}.$$

This indicates that allele A of locus 1 is on the same chromosome as allele A of locus 2, allele B of locus 1 is on the same chromosome as allele B of locus 2, etc. Alternatively we could have a genotype matrix, for example,

$$\begin{pmatrix} ABCD \\ BACD \end{pmatrix}.$$

In this case allele A of locus 1 is on the same chromosome as allele B of locus 2, etc. If the phase is uncertain, alleles will be enclosed in parentheses. For brevity in the text of this article, two-locus genotypes of known phase are also written using a slash to separate the chromosomes, so that the above genotypes would be written as $AA/BB/CC/DD$ and $AB/BA/CC/DD$, respectively.

We define P_1 and P_2 to be the phenotypes of the two parents, *i.e.*, their gel band patterns at the marker loci. P_i ($i = 1, 2$) can be denoted by a $m \times 8$ matrix, each of whose elements may take a value of 1 indicating presence of a band at the corresponding gel position or 0 indicating absence of a band. These matrices carry no information about phase. The jth rows of G_i and P_i correspond to locus M_j. Let O_{M_i} be the $n \times 8$ matrix of phenotypes of the n offspring at the marker locus M_i.

In general, there is no simple one-to-one relationship between the phenotype and the genotype of markers scored in tetraploid individuals. There are two reasons for this. First, a multiple dosage of an allele cannot be distinguished from a single dosage on the basis of the gel band pattern. Second, some alleles may not be revealed as the presence of a corresponding gel band, *i.e.*, the null alleles. Table 1 summarizes the relationship between genotype and phenotype at a marker locus in which all possible cases of null alleles and multiple dosages of identical alleles are taken into account. It can be seen from Table 1 that there may be four, six, four, or one corresponding genotype(s) if the parental phenotype shows one, two, three, or four bands. An individual genotype can be uniquely inferred from its phenotype if and only if the individual carries four different alleles and these alleles are also observed as four distinct bands.

TABLE 1

The relationship between marker phenotypes and genotypes at a single locus for an individual

Band patterns	(Phenotype)	Corresponding genotypes
One band	—	{$AOOO$}, {$AAOO$}, {$AAAO$}, {$AAAA$}
Two bands	— —	{$ABOO$}, {$AABO$}, {$ABBO$}, {$AABB$}, {$ABBB$}, {$AAAB$}
Three bands	— — —	{$ABCO$}, {$ABCC$}, {$ABBC$}, {$AABC$}
Four bands	— — — —	{$ABCD$}

Different letters represent different alleles and O denotes the null allele.

Luo *et al.* (2000) recently developed a method for predicting the probability distribution of genotypes of a pair of parents at a codominant (for example, RFLPs, microsatellites) or dominant (for example, AFLPs, RAPDs) marker locus on the basis of their and their progeny's phenotypes scored at that locus. This approach infers the number of possible configurations of the parental genotypes with the corresponding probabilities, conditional on the parental and offspring phenotypes. For each of the predicted parental genotypic configurations, the expected number of offspring phenotypes and their frequencies can be calculated and compared to the observed frequencies. Results from a simulation study and analysis of experimental data showed that in many circumstances both the parental genotypes can be correctly identified with a probability of nearly 1. A tetrasomic linkage analysis can then be carried out using the most probable parental genotype, or using each of a set of possible parental genotypes in turn if more than one genotype is consistent with all the phenotypic data. This is illustrated in the following analyses of data from simulation and experimental studies.

The steps of the linkage analysis are (i) the prediction of the parental genotype(s) that is consistent with the parental and offspring phenotype data using the method described in Luo *et al.* (2000); (ii) the detection of linkage between pairs of marker loci and their partition into linkage groups; (iii) the estimation of linkage phase, recombination frequency, and LOD score for pairs of markers within each linkage group; and (iv) the ordering of markers within each linkage group. The power to detect linkage and the variance of the estimates of the recombination frequency are shown to vary considerably with parental configuration and phase, and this will be examined.

Test for independent segregation of loci: The first step of the linkage analysis is to test whether pairs of loci are segregating independently. We propose that this may be investigated for each pair of markers by representing their joint segregation in a two-way contingency table and testing for independent segregation, as discussed by various authors (*e.g.*, Maliepaard *et al.* 1997) for diploid crosses. Let n_{ij} be the observed number of progeny with the ith ($i = 1, 2, \ldots, I$) marker phenotype at the first locus and the jth($j = 1, 2, \ldots, J$) marker phenotype at the second locus. The expected number under independent segregation is $e_{ij} = n_{i\cdot}n_{\cdot j}/n$, where $n_{i\cdot} = \Sigma_{j=1}^{J} n_{ij}$ and $n_{\cdot j} = \Sigma_{i=1}^{I} n_{ij}$. The observed and expected numbers may be compared by Pearson's chi-square statistic,

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - e_{ij})^2}{e_{ij}}. \qquad (1)$$

Other possible test statistics are the likelihood-ratio test

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\frac{n_{ij}}{e_{ij}} \qquad (2)$$

or the Cressie-Read family of power divergence statistics

$$C(\lambda) = \frac{2}{\lambda(\lambda + 1)}\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left[\left(\frac{n_{ij}}{e_{ij}}\right)^{\lambda} - 1\right]. \qquad (3)$$

These statistics have an asymptotic chi-square distribution with d.f. $= (I - 1)(J - 1)$. However, in this application the contingency tables may be sparse, as the number of cells can be as large as $36^2 = 1296$, and so the asymptotic distribution cannot be assumed without investigation. Table 2 compares the percentage points for the distribution of χ^2, G^2, and $C(\lambda)$ with $\lambda = \frac{2}{3}$ [as recommended by Cressie and Read (1984) for sparse tables] for 500 simulations of the configuration *AA*/*BB*/*CC*/*DD* × *EE*/*FF*/*GG*/*HH*, with the two loci segregating independently and 200 offspring. The percentage points for Pearson's chi-square statistic are closest to the true values, but the other two have lower percentage points. The three distributions were compared for several other configurations, but Pearson's chi-square statistic always had percentage points closest to the true distribution.

The power of Pearson's chi-square test to detect linkage was examined for 100 simulations of each of a range of configurations, linkage phases, and true recombination frequencies. For true recombination frequencies $r \leq 0.2$, the power was generally 100% (*i.e.*, the hypothesis of independent segregation was always rejected) for a significance level $\alpha = 0.01$ and >90% for $r \leq 0.3$. The exceptions to this were configurations with alleles restricted to simplex repulsion or duplex mixed configurations; *e.g.*, for cross *AB*/*AA*/*BA*/*BB* × *CC*/*CD*/*DD*/*DC*, with all alleles in duplex mixed configurations, and a true recombination frequency of 0.2, independent segregation was rejected for 3/100 simulations. When the markers were genuinely unlinked, the rejection rate for a significance level $\alpha = 0.05$ was found to be close to 5% for all configurations examined.

Partition of loci into linkage groups: Cluster analysis is a suitable technique to partition the marker loci into linkage groups, so that a marker segregates independently of markers in different linkage groups and shows a significant association with at least some of the other markers within its linkage group. The above test statistics depend on the number of marker phenotypes at each locus, but the significance level of the test for independent segregation is comparable for all pairs and could be regarded as a distance between loci. Although it ranges from 0 for the most tightly linked loci to 1, the range (0, 0.05) is of most interest for indicating pairs of loci that are likely to be linked. We therefore prefer to transform the significance level, say s, to a measure of distance that gives more discrimination between the distances of most interest. The transformation $d = 1 - 10^{-2s}$, which maps the range of the significance level (0, 0.05) to the range of the distance measure (0, 0.21), was used here, although many alternative transformations are possible. Different clustering methods will give slightly different dendrograms: the nearest-neighbor cluster analysis adds a marker to a cluster according to its distance to the closest marker in the cluster, but can combine large groups on the strength of one marker from each subgroup. We prefer to compare the dendrogram from nearest-neighbor cluster analysis with that from average linkage cluster analysis to avoid such "chaining." Inspection of the clustering at distances corresponding to different levels of significance will indicate how the marker loci should be partitioned into

TABLE 2

Percentage points for the distribution of 500 replicates of test statistics for independent segregation of two loci with parental genotypes *AA*/*BB*/*CC*/*DD* × *EE*/*FF*/*GG*/*HH*

Percentage point	True	χ^2	G^2	$C(\lambda = \frac{2}{3})$
0.25	1191	1177	703	892
0.50	1225	1213	715	915
0.75	1257	1249	726	934
0.95	1304	1295	740	959

"True" represents the true percentage points of a chi-square distribution with 1225 d.f.

linkage groups. In practice, the criterion for partitioning the dendrogram into different linkage groups can be determined as the distance measure by which significant linkage is inferred. However, the Bonferroni correction for the overall significance level may be necessary to take the multiple linkage tests into account. The calculation of recombination frequencies and LOD scores then proceeds for each linkage group in turn.

Calculation of segregation probabilities: One of the major difficulties in linkage analysis with tetraploid species is to calculate the conditional distribution of the offspring genotypes, and hence phenotypes, at two linked loci for any given pair of parental genotypes. This involves consideration of a large number of segregation and recombination events. In this section, a general computer-based algorithm is described to compute the probability distribution.

For simplicity but without loss of generality, we use A and B for two loci in this section and subscripts to represent the alleles. Consider a parental genotype $A_iB_i/A_jB_j/A_kB_k/A_lB_l$. During gametogenesis of the individual, three equally likely pairs of bivalents can be generated, *i.e.*, $A_iB_i/A_jB_j//A_kB_k/A_lB_l$, $A_iB_i/A_kB_k//A_jB_j/A_lB_l$, and $A_iB_i/A_lB_l//A_jB_j/A_kB_k$, where // is used to distinguish paired homologous chromosomes. The gametes created from each of these pairs of bivalents can be sorted into three classes: (i) nonrecombinants, $A_\xi B_\xi A_\eta B_\eta$($\xi \neq \eta$; ξ and η may be i, j, k, or l), four gametic genotypes, each of which has a frequency of $(1 - r)^2/4$; (ii) single recombinants $A_\xi B_\eta A_\gamma B_\gamma$($\xi \neq \eta \neq \gamma$; ξ, η, or γ may be i, j, k, or l), eight gametic genotypes, each with a frequency of $r(1 - r)/4$; (iii) double recombinants $A_\xi B_\eta A_\gamma B_\zeta$ ($\xi \neq \eta \neq \gamma \neq \zeta$; ξ, η, γ, or ζ may be i, j, k or l), four gametic genotypes, each with a frequency of $r^2/4$. Thus, when the three possible pairs of bivalents are considered, a general form for frequency of the gametic genotype i can be written as

$$g_i = \frac{x_{i0}}{12}(1 - r)^2 + \frac{x_{i1}}{12}r(1 - r) + \frac{x_{i2}}{12}r^2, \qquad (4)$$

where x_{i0}, x_{i1}, and x_{i2} are numbers of the nonrecombinants, single recombinants, and double recombinants, respectively, within the ith gametic genotype class. With random union between all possible gametes generated from two parents and sorting the zygotes according to their genotype, a general formula for the frequency of zygote genotype i may be expressed as

$$\begin{aligned} h_i &= \frac{1}{144}[y_{i0}(1 - r)^4 + y_{i1}r(1 - r)^3 + y_{i2}r^2(1 - r)^2 \\ &\quad + y_{i3}r^3(1 - r) + y_{i4}r^4] \\ &= \frac{1}{144}\sum_{j=0}^{4} y_{ij}r^j(1 - r)^{4-j}, \end{aligned}$$

where y_{ij} is the number of zygotes with j recombinations within the ith zygote genotype.

To evaluate the coefficients y_{ij} manually is obviously very tedious. A computer algorithm was developed to calculate the offspring's genotypic distribution for any given pair of tetraploid parental genotypes. The computer subroutine outputs the number of all possible distinct offspring genotypes k and $\{y_{ij}\}$ ($i = 1, 2, \ldots, k$) from the two parental genotypes. For example, if two parental genotypes are *AA/BB/BB/OB* and *CA/DA/EC/EO*, there are a total of 225 possible genotypes in their offspring. Many of these offspring genotypes correspond to the same phenotype. Thus, the phenotypic distribution of the offspring can be readily derived by combining the probabilities of those genotypes that result in the same phenotype, so that the general formula for the probability of zygote phenotype i is

$$f_i = \sum_{g\in i} h_g = \frac{1}{144}\sum_{g\in i}\sum_{j=0}^{4} y_{gj}r^j(1 - r)^{4-j}. \qquad (5)$$

In the above equation, $\Sigma_{g\in i}h_g$ indicates the sum over the frequencies of all those genotypes g that correspond to the same phenotype i. For instance, the 225 offspring genotypes in the above example are classified into 36 distinct phenotypes when the marker genes are assumed to be codominant, and these are illustrated in Table 3. It can be seen that the frequency of the first phenotype is $f_1 = [8(1 - r)^4 + 32r(1 - r)^3 + 40r^2(1 - r)^2 + 24r^3(1 - r) + 8r^4]/144 = (1 - r^2 + r^3)/18$.

Maximum-likelihood estimate of r: If the parental genotypes and their linkage phase are known, the joint expected phenotypic distribution of their offspring can be derived using the method suggested above. The corresponding observed offspring phenotypes at the marker loci can be recognized as a random sample from a multinomial distribution with probabilities f_i ($i = 1, 2, \ldots, k$) and sample size $n = \Sigma_{i=1}^{k}n_i$, where k is the number of possible phenotypes and n_i is the observed number of offspring in the ith phenotype class. Thus, the log-likelihood of the recombination frequency, r, given the observed data at loci M_i and M_j, is given by

$$\begin{aligned} L\{r|O_{M_i}, O_{M_j}, G_1, G_2\} &= \ln\left\{\binom{n}{n_1, n_2, \ldots, n_k} f_1^{n_1}f_2^{n_2}\ldots f_k^{n_k}\right\} \\ &= C + \sum_{i=1}^{k} n_i\ln(f_i), \end{aligned} \qquad (6)$$

where f_i ($i = 1, 2, \ldots, k$) are functions of r and given by Equation 5.

The maximum-likelihood estimate (MLE) of the recombination frequency r may be obtained by solving

$$\frac{d}{dr}L\{r|O_{M_i}, O_{M_j}, G_1, G_2\} = \sum_{i=1}^{k}\frac{n_i}{f_i}\frac{d}{dr}(f_i) = 0. \qquad (7)$$

Only in a limited number of cases can the likelihood equation be solved analytically because the equation is usually a polynomial with a power ≥5. An iterative solution may be obtained, however, using the expectation-maximization (EM) algorithm (Dempster *et al.* 1977). Maliepaard *et al.* (1997) applied the EM algorithm to give a general formulation for all possible genetic

TABLE 3

Phenotypic distribution of a full-sib family from crossing two autotetraploid genotypes *AA/BB/BB/OB* and *CA/DA/EC/EO*

Class	Phenotype at locus 1	Phenotype at locus 2	y_{i0}	y_{i1}	y_{i2}	y_{i3}	y_{i4}
1	1 1 1 0 1 0 0 0	1 1 1 0 0 0 0 0	8	32	40	24	8
2	1 1 0 1 1 0 0 0	1 1 1 0 0 0 0 0	8	32	40	24	8
3	1 1 1 0 1 0 0 0	1 1 0 0 0 0 0 0	8	32	48	32	8
4	1 1 0 1 1 0 0 0	1 1 0 0 0 0 0 0	8	32	48	32	8
5	1 1 1 1 0 0 0 0	1 1 0 0 0 0 0 0	8	24	24	8	0
6	1 1 0 0 1 0 0 0	1 1 1 0 0 0 0 0	8	16	16	8	0
7	1 1 1 1 0 0 0 0	1 1 1 0 0 0 0 0	0	8	24	16	0
8	1 1 0 0 1 0 0 0	1 1 0 0 0 0 0 0	0	8	24	24	8
9	0 1 1 0 1 0 0 0	1 1 1 0 0 0 0 0	12	36	60	48	12
10	0 1 0 1 1 0 0 0	1 1 1 0 0 0 0 0	12	36	60	48	12
11	0 1 1 0 1 0 0 0	1 1 0 0 0 0 0 0	12	48	72	48	12
12	0 1 0 1 1 0 0 0	1 1 0 0 0 0 0 0	12	48	72	48	12
13	0 1 1 1 0 0 0 0	1 1 0 0 0 0 0 0	12	36	36	12	0
14	0 1 0 0 1 0 0 0	0 1 1 0 0 0 0 0	12	12	0	0	0
15	0 1 1 1 0 0 0 0	1 1 1 0 0 0 0 0	0	12	24	24	12
16	0 1 0 1 1 0 0 0	0 1 1 0 0 0 0 0	0	12	12	0	0
17	0 1 0 0 1 0 0 0	1 1 1 0 0 0 0 0	0	24	36	12	0
18	0 1 1 0 1 0 0 0	0 1 1 0 0 0 0 0	0	12	12	0	0
19	0 1 0 0 1 0 0 0	1 1 0 0 0 0 0 0	0	12	36	36	12
20	0 1 1 1 0 0 0 0	0 1 1 0 0 0 0 0	0	0	12	12	0
21	1 0 1 0 1 0 0 0	1 1 1 0 0 0 0 0	4	16	20	12	4
22	1 0 0 1 1 0 0 0	1 1 1 0 0 0 0 0	4	16	20	12	4
23	1 0 1 0 1 0 0 0	1 1 0 0 0 0 0 0	4	16	24	16	4
24	1 0 0 1 1 0 0 0	1 1 0 0 0 0 0 0	4	16	24	16	4
25	1 0 1 1 0 0 0 0	1 1 0 0 0 0 0 0	4	12	12	4	0
26	1 0 0 0 1 0 0 0	1 1 1 0 0 0 0 0	4	8	8	4	0
27	1 0 1 1 0 0 0 0	1 1 1 0 0 0 0 0	0	4	12	8	0
28	1 0 0 0 1 0 0 0	1 1 0 0 0 0 0 0	0	4	12	12	4
29	1 1 0 0 1 0 0 0	0 1 1 0 0 0 0 0	0	8	8	0	0
30	1 1 0 1 1 0 0 0	0 1 1 0 0 0 0 0	0	0	8	8	0
31	1 1 1 0 1 0 0 0	0 1 1 0 0 0 0 0	0	0	8	8	0
32	1 1 1 1 0 0 0 0	0 1 1 0 0 0 0 0	0	0	0	8	8
33	1 0 0 0 1 0 0 0	0 1 1 0 0 0 0 0	0	4	4	0	0
34	1 0 0 1 1 0 0 0	0 1 1 0 0 0 0 0	0	0	4	4	0
35	1 0 1 0 1 0 0 0	0 1 1 0 0 0 0 0	0	0	4	4	0
36	1 0 1 1 0 0 0 0	0 1 1 0 0 0 0 0	0	0	0	4	4

configurations in a cross between two outbreeding diploid plant species, and here we modify their approach for the autotetraploid case.

In Equation 5, define $z_{ij} = \Sigma_{g\in i} y_{gj} r^j (1 - r)^{4-j}/144$, so that the probability of phenotype i is $f_i = \Sigma_{j=0}^{4} z_{ij}$. Substituting this into Equation 7, we obtain

$$\begin{aligned}\frac{d}{dr}L\{r|O_{M_i}, O_{M_j}, G_1, G_2\} &= \sum_{i=1}^{k}\frac{n_i}{f_i}\sum_{j=0}^{4}\frac{d}{dr}(z_{ij})\\ &= \sum_{i=1}^{k} n_i \sum_{j=0}^{4}\frac{z_{ij}}{f_i}\frac{d}{dr}(\ln(z_{ij}))\\ &= \sum_{i=1}^{k} n_i \sum_{j=0}^{4}\frac{z_{ij}}{f_i}\frac{j-4r}{r(1-r)}. \qquad (8)\end{aligned}$$

So the derivative of the likelihood is equal to zero when

$$r = \frac{1}{4n}\sum_{i=1}^{k} n_i \sum_{j=0}^{4}\frac{jz_{ij}}{f_i}. \qquad (9)$$

From an initial estimate of r, we can calculate z_{ij}/f_i, the expected proportion of individuals of phenotype i with j recombinants (the expectation step of the EM algorithm), and substitute this into Equation 9 to give an updated estimate of r (the maximization step). The algorithm is iterated until the sequence of estimates of r converges.

Estimation of parental pairwise linkage phases: In the above analyses, it was assumed that the parental genotypes and their linkage phases were known. In practice, only the parental and offspring phenotypes are observable. As pointed out in the *Model and notation* section, Luo *et al.* (2000) calculate the genotypic distribution for any pair of tetraploid parents at a single dominant or codominant marker locus using data on the marker phenotypes scored on the parents and their offspring. However, the method does not provide information about the linkage phases of the alleles the parents carry at different loci. Knowledge about the linkage phase of the parental genotypes is not only required in the linkage analysis, but it is also important in using the map information in locating QTL (*e.g.*, Lander and Botstein 1989) or optimizing schemes of marker-assisted selection for quantitative traits (Luo *et al.* 1997).

The number of possible different linkage phases depends on the number of distinct alleles at each locus and increases exponentially with the number of loci under consideration. We therefore consider here the phase for each pair of linked loci and use these as building blocks to estimate the multilocus linkage phase. In a two-locus system of tetrasomic inheritance, an individual genotype may have a maximum of 4 × 3 × 2 = 24 distinct linkage phases, and for a pair of individuals there may be a maximum of 24 × 24 = 576 distinct linkage phase configurations. A Fortran-90 computer subroutine was developed to work out all possible linkage phase configurations for any given pair of parental genotypes G_1 and G_2 at any two loci i and j. Let S_1 and S_2 be possible two-locus linkage phases for parents 1 and 2, respectively. The likelihood of r, S_1, and S_2, given the observed phenotypic data O_{M_i} and O_{M_j} at the loci, may be written as

$$l_p[r, S_1, S_2|O_{M_i}, O_{M_j}, G_1, G_2] = \Pr\{O_{M_i}, O_{M_j}|r, S_1, S_2\}$$
$$= \binom{n}{n_1, n_2, \ldots, n_k} f_1^{n_1} f_2^{n_2} \ldots f_k^{n_k}, \quad (10)$$

where the expected frequency of the ith offspring phenotype, f_i ($i = 1, 2, \ldots, k$), is calculated for given r, S_1, and S_2 as demonstrated in Equation 5. As discussed above, the EM algorithm can be used to maximize the likelihood function of Equation 10 for every possible configuration of S_1 and S_2. The configuration $\hat{S}_1$, $\hat{S}_2$ for which this maximum is the highest is taken as the maximum-likelihood estimate of the parental genotypic linkage phase and the corresponding value of $\hat{r}$ is the maximum-likelihood estimate of r. The LOD score for each pair of marker loci is calculated as

$$\mathrm{LOD} = \log_{10} \frac{l_p[\hat{r}, \hat{S}_1, \hat{S}_2|O_{M_i}, O_{M_j}, G_1, G_2]}{l_p[0.5, \hat{S}_1, \hat{S}_2|, O_{M_i}O_{M_j}, G_1, G_2]}. \quad (11)$$

Ordering the markers: The above analyses give the maximum-likelihood estimate of the recombination frequency and the linkage phase for each pair of markers in a linkage group. This information can be used to order the markers in linkage groups and to calculate map distances between them. One possible approach, the least-squares method for estimation of multilocus map distances as implemented in the JoinMap linkage software (Stam and Van Ooijen 1995), was examined by Hackett *et al.* (1998) in a simulation study of dominant markers in a tetraploid population. They concluded that the reconstructed marker order and map distance using the JoinMap analysis of simulation data was in good agreement with the simulated ones. The same method was used here.

Estimation of parental multilocus linkage phases: Once the markers have been ordered, we need to reconstruct the phase of the complete linkage group. Prediction of the multilocus linkage phase in tetrasomic linkage analysis is not feasible: there are a huge number of configurations of possible phases and no appropriate theory of multilocus linkage analysis for tetraploid species. Here we propose an intuitive algorithm to predict the multilocus parental linkage phase in the tetrasomic linkage analysis, on the basis of the range of likelihood values of the alternative linkage phases obtained in the above two-locus analysis. Let d_{ij} be the difference in the log-likelihood value between the most likely and the second most likely linkage phases predicted for the marker loci i and j on a linkage group. The phase of the marker pair with the largest log-likelihood difference d_{ij} is reconstructed first, and further markers are then placed relative to this pair, placing markers with large d_{ij} before those with smaller d_{ij}. There may be a contradiction between the phase of two markers estimated directly and the phase estimated when each of the pair is referred to a third marker; we reject an overall configuration with such contradictions for a pair with large d_{ij}, but accept the overall configuration if d_{ij} is close to zero.

INFORMATION AND POWER OF THE MAXIMUM-LIKELIHOOD ESTIMATION

The information of the maximum-likelihood estimate of the recombination frequency r is given by

$$I(r) = -E\left[\frac{d^2}{dr^2} L_p[r, S_1, S_2|O_{M_i}, O_{M_j}, G_1, G_2]\right]$$
$$= -E\left[\sum_{i=1}^{k} n_i\left[\frac{1}{f_i}\frac{d^2}{dr^2}(f_i) - \frac{1}{f_i^2}\left(\frac{d}{dr}(f_i)\right)^2\right]\right], \quad (12)$$

where E denotes expectation. It can be shown that the expectation of the first term on the right-hand side of this expression is equal to zero. Substituting for the second term, we obtain

$$I(r) = E\left[\sum_{i=1}^{k} \frac{n_i}{f_i^2}\left(\frac{d}{dr}(f_i)\right)^2\right]$$
$$= E\left[\sum_{i=1}^{k} \frac{n_i}{f_i^2}\left(\sum_{j=0}^{4}\frac{d}{dr}(z_{ij})\right)^2\right]$$
$$= \frac{n}{r^2(1-r)^2}\left[\sum_{i=1}^{k}\frac{1}{f_i}\left(\sum_{j=0}^{4} jz_{ij}\right)^2 - 16r^2\right]. \quad (13)$$

The details of the derivation of this equation are given in the appendix.

Hackett *et al.* (1998) demonstrated that the simplex coupling linkage phase was the most informative for estimating recombination frequency among dominant marker configurations. For this the information is $n/r(1-r)$ and the information content of other configurations is examined relative to this by means of the relative information

$$\mathrm{RI}(r) = \frac{I(r)}{n/r(1-r)} = \frac{1}{r(1-r)}\left[\sum_{i=1}^{k}\frac{1}{f_i}\left(\sum_{j=0}^{4} j z_{ij}\right)^2 - 16r^2\right]. \tag{14}$$

We can also examine the power of the likelihood-ratio test. The likelihood-ratio test statistic is given by

$$G^2 = 2\{\ln\left[l_p[\hat{r},\ \hat{S}_1,\ \hat{S}_2 | O_{M_i},\ O_{M_j},\ G_1,\ G_2]\right] - \ln\left[l_p[r = 0.5,\ \hat{S}_1,\ \hat{S}_2 | O_{M_i},\ O_{M_j},\ G_1,\ G_2]\right]\}. \tag{15}$$

It has been shown by Agresti (1990, pp. 98, 241) that G^2 has an approximate large-sample noncentral chi-square distribution with 1 d.f. and the noncentral parameter in the present context is

$$\lambda = 2n\sum_{i=1}^{k} f_i(r)\ln\left[\frac{f_i(r)}{f_i(0.5)}\right]. \tag{16}$$

Thus, the statistical power for the linkage test at a given significance level α is given by the probability

$$\beta_L = \Pr\{\chi^2_{1,\lambda} > \chi^2_1(\alpha)\}, \tag{17}$$

where $\chi^2_{1,\lambda}$ represents a random variable with a noncentral chi-square distribution with 1 d.f. and the noncentrality parameter λ, and $\chi^2_1(\alpha)$ is the $1 - \alpha$ percentile of a central chi-square distribution, also with 1 d.f.

For two parents, there are 128 configurations at a single locus where the parents share one or more alleles, which are informative about recombination in both parents. This count does not include permutations of the parents; *i.e.*, *AAOO* × *AOOO* and *AOOO* × *AAOO* are considered as the same configuration. To consider all pairs of such loci, and to allow for the different phases, would give a very large number of configurations. We therefore examined the information and power of the likelihood-ratio test for each configuration when linked to a locus with eight alleles, *ABCD* × *EFGH*. The most informative configurations are those with seven or eight alleles: *AA/BB/CC/DD* × *EE/FF/GG/HH* and *AA/BB/CC/DD* × *EA/FE/GF/HG*, which are four times as informative as the simplex coupling configuration for all values of the recombination frequency. For many configurations, the relative information varies with the recombination frequency. At a recombination frequency of 0.2, 20 of the configurations examined were less informative than simplex coupling: these configurations were characterized by a small number of alleles occurring as simplex or duplex in each parent. The least informative configuration was *AA/BA/CO/DO* × *EA/FA/GO/HO*, with a relative information of 0.14. There was a strong linear relationship between the information and the noncentrality of the likelihood-ratio test, for example, a correlation of 0.996 using a recombination frequency of 0.2. For a recombination frequency of 0.2 and a population of 200 offspring, the power of the likelihood-ratio test was >0.9 for all configurations except the least informative *AA/BA/CO/DO* × *EA/FA/GO/HO*, although the power decreases with decreasing population size or increasing marker separation.

When the two parents do not share any alleles, the information can be calculated for each parent separately and then summed. The most informative configuration for a single parent is *AA/BB/CC/DD*, which is twice as informative as the simplex coupling configuration for all values of the recombination frequency. The relationship between the information and the noncentrality is the same as for two parents with shared alleles. The least informative configurations are some of those with a single informative allele: duplex-duplex mixed (*AA/AO/OA/AA*, relative information = 0.04), simplex repulsion (*AO/OA/OO/OO*, relative information = 0.07), and duplex-duplex repulsion (*AO/AO/OA/OA*, relative information = 0.11). Some configurations with two informative alleles also have very low information, for example, *AO/AB/BA/OA*, where the two duplex alleles at each locus are in repulsion and so are the two simplex alleles. The relationship between the relative information and the recombination frequency is illustrated for a range of configurations in Figure 1.

As the information depends on the configurations of both loci and on their phase, it is difficult to exclude any single-locus configurations as uninformative. The more alleles at a locus, the more informative it is likely to be, especially if these loci are present in a simplex configuration. A locus with an allele that occurs in both parents is likely to have a low information content, unless we are considering a configuration such as *AA/OO/OO/OO* × *AA/OO/OO/OO*, with single-dose alleles in coupling in both parents. The configurations with low power for detecting nonindependent segregation by Pearson's chi-square test also had low information and low power in the likelihood-ratio test.

SIMULATION STUDY

To validate the theoretical analyses represented above and to investigate their statistical properties, we conducted a simulation study using the method developed above.

Simulation model: Computer programs were developed to simulate meiosis in a tetraploid individual with any genotype at the simulated marker loci, random pairing of four homologous chromosomes to give two bivalents (*i.e.*, no double reduction), random sampling of gametes from meiosis, random union of gametes randomly sampled from the gamete pool, and generation of the phenotype from any given individual genotype. In a single meiosis, the "random walk" procedure sug-

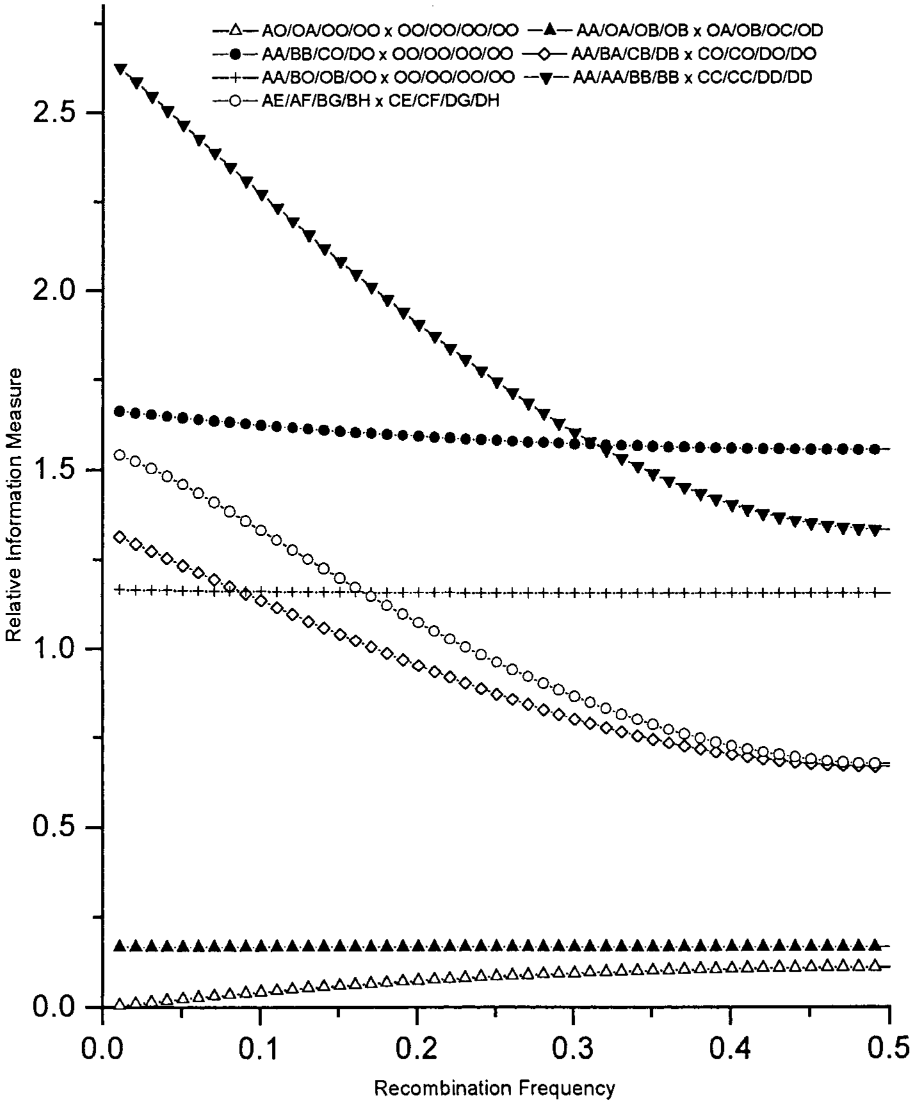


FIGURE 1.—Relative information about the recombination frequency for different parental genotype configurations.

gested by CROSBY (1973) was extended to simulate genetic recombination between linked loci. Chiasmata interference, sexual differentiation in recombination frequency, and segregation distortion were assumed to be absent in the simulation model.

A full-sib family was simulated by crossing two tetraploid parental lines. Twenty-two codominant marker loci were generated, 10 linked on the first chromosome, 5 on each of the second and third chromosomes, and 2 isolated loci that were independent of the rest. The simulated parental genotypes at each of the marker loci were determined by sampling independently from six possible alleles whose population frequencies were assumed to be 0.3 (allele *A*), 0.2 (allele *B*), 0.2 (allele *C*), 0.1 (allele *D*), 0.1 (allele *E*), and 0.1 (null allele *O*), respectively. Loci with more than six alleles were not simulated, as these appear to be rare in practice (R. C. MEYER, personal communication). The main purpose for choosing parental genotypes in such a way is to test the theory and method on a general basis. The parental genotypes at these marker loci and the recombination frequencies between the adjacent loci are shown in Table 4. It should be noted that the alleles listed in the same column for loci on the same chromosome have the same linkage phase. The phenotypes of the two parents and 200 offspring (a realistic number for actual experiments) were scored at all 22 marker loci. To elucidate statistical properties, some pairs of these loci were studied in 100 repeated simulation trials.

Analysis of the simulated data: The genotypes of the two parents were predicted for each of the loci using the method proposed by LUO *et al.* (2000), on the basis of the phenotypes of the parents and their offspring. The predicted parental genotypes are tabulated in Table 4 together with the corresponding probabilities. It can be seen that the parental genotypes at 18 of the 22 marker loci were diagnosed correctly with a prediction probability of nearly 1.0. However, there were two almost equally likely parental genotypes predicted for the marker loci L_2, L_5, L_{20}, and L_{22}. For locus L_2 the parental

TABLE 4

The simulated parental genotypes (G_1 and G_2), their corresponding phenotypes (P_1 and P_2), and the most likely parental genotypes ($\hat{G}_1 \times \hat{G}_2$) predicted at 22 simulated marker loci

Loci	Chr	r	G_1	G_2	P_1	P_2	$\hat{G}_1 \times \hat{G}_2$ (prob)
L_1	1	0.00	*CABB*	*DCEO*	11100000	00111000	*ABBC* × *CDEO* (1.0000)
L_2	1	0.10	*CABA*	*BCCA*	11100000	11100000	*AABC* × *ABCC* (0.4999) or *ABCC* × *AABC* (0.4999)
L_3	1	0.10	*BCAE*	*ACAB*	11101000	11100000	*ABCE* × *AABC* (0.9999)
L_4	1	0.05	*OBCA*	*AABD*	11100000	11010000	*ABCO* × *AABD* (1.0000)
L_5	1	0.10	*AAAO*	*CDCC*	10000000	00110000	*AAAO* × *CCCD* (0.5000) or *AAAA* × *CCCD* (0.5000)
L_6	1	0.05	*DOAE*	*ABAB*	10011000	11000000	*ADEO* × *AABB* (0.9999)
L_7	1	0.10	*BOAA*	*DABB*	11000000	11010000	*AABO* × *ABBD* (0.9989)
L_8	1	0.05	*BBDB*	*ABAD*	01010000	11010000	*BBBD* × *AABD* (0.9999)
L_9	1	0.10	*DDBE*	*BBAD*	01011000	11010000	*BDDE* × *ABBD* (0.9999)
L_{10}	1	0.05	*AEDE*	*DACA*	10011000	10110000	*ADEE* × *AACD* (1.0000)
L_{11}	2	0.50	*AACB*	*ACDD*	11100000	10110000	*AABC* × *ACDD* (1.0000)
L_{12}	3	0.50	*ACBB*	*ACAA*	11100000	10100000	*ABBC* × *AAAC* (0.9999)
L_{13}	4	0.50	*AOCA*	*AEEB*	10100000	11001000	*AACO* × *ABEE* (0.9999)
L_{14}	4	0.10	*BEBD*	*ABCC*	01011000	11100000	*BBDE* × *ABCC* (1.0000)
L_{15}	4	0.10	*BOBA*	*AOCA*	11000000	10100000	*ABBO* × *AACO* (0.9998)
L_{16}	4	0.10	*DCCA*	*BOAD*	10110000	11010000	*ACCD* × *ABDO* (1.0000)
L_{17}	4	0.10	*OBEA*	*BCDC*	11001000	01110000	*ABEO* × *BCCD* (1.0000)
L_{18}	5	0.50	*DCBC*	*CDDO*	01110000	00110000	*BCCD* × *CDDO* (0.9881)
L_{19}	5	0.20	*AOOE*	*CCOB*	10001000	01100000	*AOOE* × *BCCO* (1.0000)
L_{20}	5	0.20	*AAAC*	*ADAC*	10100000	10110000	*AAAC* × *AACD* (0.4988) or *AAAC* × *ACDO* (0.4988)
L_{21}	5	0.20	*ODBB*	*BCAE*	01010000	11101000	*BBDO* × *ABCE* (1.0000)
L_{22}	5	0.20	*AADA*	*ABAC*	11010000	11100000	*AAAD* × *AABC* (0.4999) or *AAAD* × *ABCO* (0.4900)

Chr, the linkage group number, and r, the recombination frequencies between adjacent loci, are the most likely predicted parental genotypes.

phenotypes are the same (1110000), but the most likely parental genotypes are different (*AABC* and *ABCC*) and it is not possible at this stage to tell which parent has which genotype. Both genotypes at this locus were used in the linkage analysis. For the other three loci, allele *A* is present for all offspring, and this is consistent with more than one configuration with multiple dosages of *A*. The dosages of the informative alleles are the same for the two possible configurations for L_5, L_{20}, and L_{22}, and so estimates of recombination frequencies are the same for the two configurations.

Pearson's chi-square tests of independence were performed for all possible pairs of these marker loci using the test statistic given in Equation 1. Figure 2 displays the significance probabilities, transformed to distances as described previously, as dendrograms calculated using nearest-neighbor cluster analysis and average linkage cluster analysis. The nearest-neighbor analysis shows the three clusters (loci L_1–L_{10}, L_{13}–L_{17}, and L_{18}–L_{22}) have each grouped at a distance of zero. Loci L_{11} and L_{12} remained isolated until the distance exceeded 0.13. However, the three linkage groups also merge at a very small distance. Inspection of the significance levels shows that this is due to a single (spurious) significant

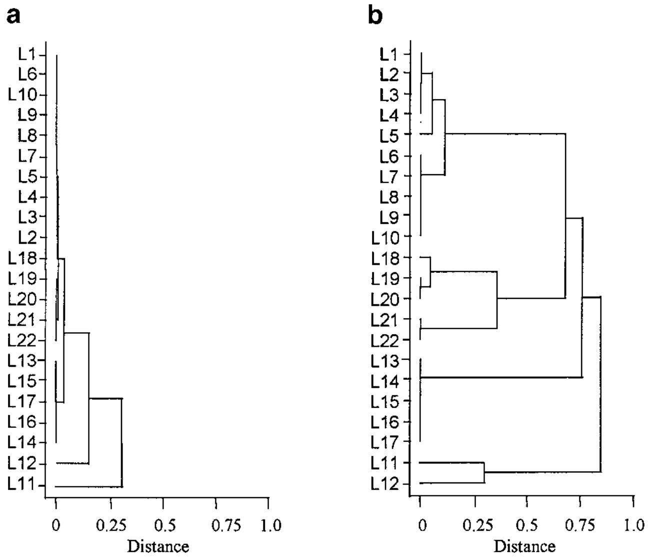


Figure 2.—Cluster analysis of the 22 simulated loci, using (a) nearest-neighbor cluster analysis and (b) average linkage cluster analysis.

TABLE 5

The maximum-likelihood estimates of pairwise recombination frequencies (the upper diagonal) and the LOD scores (the second rows of the lower diagonal) calculated for the most likely parental phases for loci L_1–L_{10}

Loci	L_1	L_2	$L_{2'}$	L_3	L_4	L_5	L_6	L_7	L_8	L_9	L_{10}
L_1	—	0.11	0.18	0.19	0.18	0.23	0.24	0.33	0.36	0.40	0.39
L_2	0.000	—	—	0.13	0.10	0.13	0.22	0.30	0.34	0.35	0.33
	31.03	—	—								
$L_{2'}$	0.000	—	—	0.22	0.13	0.19	0.27	0.29	0.32	0.37	0.36
	19.67	—	—								
L_3	0.000	0.000	0.000	—	0.04	0.08	0.15	0.31	0.18	0.33	0.35
	24.00	11.28	7.28								
L_4	0.000	0.000	0.000	0.000	—	0.03	0.09	0.16	0.21	0.28	0.28
	35.55	24.59	17.86	47.82							
L_5	0.000	0.000	0.000	0.000	0.000	—	0.02	0.18	0.11	0.22	0.37
	8.30	4.10	4.13	14.16	9.52						
L_6	0.000	0.000	0.000	0.000	0.000	0.000	—	0.11	0.09	0.24	0.26
	18.33	7.16	5.13	37.57	33.91	12.38					
L_7	0.012	0.001	0.001	0.406	0.000	0.000	0.000	—	0.04	0.16	0.20
	7.22	4.00	3.71	3.21	11.44	6.97	16.35				
L_8	0.013	0.461	0.461	0.000	0.000	0.000	0.000	0.000	—	0.17	0.20
	2.96	0.97	1.43	7.48	14.13	7.99	14.20	11.29			
L_9	0.735	0.349	0.349	0.021	0.000	0.090	0.000	0.000	0.000	—	0.06
	2.70	1.96	1.32	6.50	10.84	2.21	15.63	10.45	7.76		
L_{10}	0.418	0.000	0.000	0.013	0.000	0.410	0.000	0.000	0.000	0.000	—
	3.44	4.16	2.89	3.19	8.71	0.66	9.96	11.08	6.59	56.26	

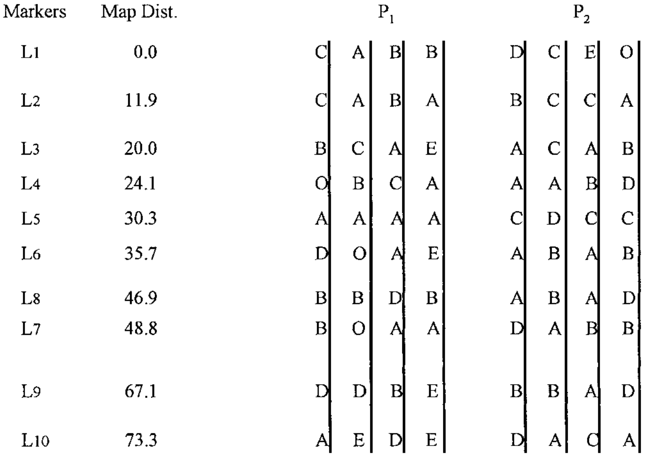
Listed in the first rows of the lower diagonal are the significance of the independent tests.

association between L_1 and L_{18}. Inspection of the average linkage cluster analysis shows that the same initial groupings form more slowly, but that locus L_{18} is clearly associated with L_{19} and L_{20}, and the average distance between locus L_{18} and loci L_1–L_{10} is large. The distantly linked group L_{18}–L_{22} finally merges at a large distance using average linkage cluster analysis, but inspection of the significance levels shows highly significant associations between 6 of the 10 pairs of this group, and we proceed assuming that they form a linkage group.

Linkage analysis was performed on all pairs of loci within each linkage group. For brevity only the results from the largest linkage group (loci L_1–L_{10}) are presented. Table 5 shows the significance of this test (the first rows of the lower diagonal), the maximum-likelihood estimates of recombination frequencies (the upper diagonal) for the most likely phase, and the corresponding LOD scores (the second rows of the lower diagonal) among the pairs of loci. It can be seen that the true parental genotype at L_2 has consistently higher LOD scores in its pairings with L_1, L_3, L_4, L_6, and L_7 (*i.e.*, all the highly significant linkages) than the other predicted parental genotype ($L_{2'}$) with the parental genotypes reversed. The estimate of the recombination frequency and LOD score were unaffected by the choice between the alternative genotypes for locus L_5. Most cases where the independence test was significant ($P <$ 0.05) corresponded to LOD scores >3, although a small number of pairs with a significant independence test (*e.g.*, L_1, L_8) had large recombination frequencies and lower LODs.

The maximum-likelihood estimates of the pairwise recombination frequencies and the LOD scores in Table 5 were used to construct a linkage map of these genetic marker loci using the JoinMap linkage software (Stam and Van Ooijen 1995), as summarized in Figure 3. The best-fitted map predicted from JoinMap indicates that loci L_1–L_{10} were joined into a correct order except that the relative simulated positions of the marker loci L_7 and L_8 were reversed. The map distances of the linkage group agreed well with the actual ones.

The linkage phases of the parental genotypes were reconstructed using the procedure described in the above analysis. Table 6 illustrates the parental linkage phases at every pair of loci with a difference $d_{ij} > 3$ in the log-likelihood between the most likely and second most likely phase. Locus L_5 does not appear in Table 6, as there was only one phase with a recombination frequency <0.5 in each case. The reconstructed linkage phases of the parental genotypes are shown in Figure 3. This reconstruction uses the most likely phase for all pairs except for four [(L_1, L_8), (L_1, L_9), (L_2, L_7), (L_6, L_9)]. For these four pairs, the largest difference in the log-likelihood between the most likely phase and the reconstructed phase was 0.56, and the difference in the estimates of the recombination frequency was always <0.01. The reconstructed phase is identical to that simulated.



Markers	Map Dist.	P_1				P_2			
L1	0.0	C	A	B	B	D	C	E	O
L2	11.9	C	A	B	A	B	C	C	A
L3	20.0	B	C	A	E	A	C	A	B
L4	24.1	O	B	C	A	A	A	B	D
L5	30.3	A	A	A	A	C	D	C	C
L6	35.7	D	O	A	E	A	B	A	B
L8	46.9	B	B	D	B	A	B	A	D
L7	48.8	B	O	A	A	D	A	B	B
L9	67.1	D	D	B	E	B	B	A	D
L10	73.3	A	E	D	E	D	A	C	A

The goodness of fit test $\chi^2 = 28.7$ with d.f.=36

Figure 3.—The best-fitted map, the estimated map distance (in centimorgans), and parental linkage phases reconstructed from the codominant marker loci L_1–L_{10} from the simulation study.

TABLE 6

The most likely parental genotypic linkage phases (S_1 and S_2) and the difference (d_{ij}) in log-likelihood value between the most likely and the second most likely linkage phases of the marker loci L_i and L_j

L_i	L_j	S_1	S_2	d_{ij}
1	2	*CC/AA/BB/BA*	*DB/CC/EC/OA*	10.18
1	3	*CB/AC/BA/BE*	*DA/CC/EA/OB*	15.19
1	4	*CO/AB/BC/BA*	*DA/CA/EB/OD*	7.72
1	6	*CD/AO/BA/BE*	*DA/CB/EA/OB*	4.32
2	3	*CB/AC/BA/AE*	*BA/CC/CA/AB*	4.37
3	4	*BO/CB/AC/EA*	*AA/CA/AB/BD*	18.72
3	6	*BD/CO/AA/EE*	*AA/CB/AA/BB*	4.19
3	8	*BB/CB/AD/EB*	*AA/CB/AA/BD*	3.51
4	6	*OD/BO/CA/AE*	*AA/AB/BA/DB*	11.67
4	8	*OB/BB/CD/AB*	*AA/AB/BA/DD*	8.87
4	9	*OD/BD/CB/AE*	*AB/AB/BA/DD*	3.95
4	10	*OA/BE/CD/AE*	*AD/AA/BC/DA*	3.49
7	8	*BB/OB/AD/AB*	*DA/AB/BA/BD*	4.32
8	9	*BD/BD/DB/BE*	*AB/BB/AA/DD*	3.65
9	10	*DA/DE/BD/EE*	*BD/BA/AC/DA*	14.64

Linkage maps of loci L_{13}–L_{17} and L_{18}–L_{22} were estimated using the same approach. In each case the order and phase were reconstructed correctly.

To investigate the reliability of the pairwise linkage phase estimation, separate simulation trials were performed. The simulated recombination frequencies were 0.05, 0.1, and 0.3 and the sample size was 200. Figure 4 illustrates the maximum-likelihood estimate of r between L_9 and L_{10} (Figure 4a), and between L_7 and L_{10} (Figure 4b), and the corresponding LOD scores calculated at all possible parental linkage phase configurations. It can be seen that the correct parental linkage phases were the most likely when the marker loci were closely linked (*i.e.*, $r \leq 0.1$), although the difference in the likelihood value between the most likely and the second most likely linkage phases reduced as the value of r increased. When the loci were loosely linked (*i.e.*, $r = 0.3$), the most likely parental linkage phase could differ from the simulated phase, but when this occurred the MLE of r at the most likely phase was always very close to that calculated at the simulated phase.

Further simulations were carried out to examine the power to detect linkage and the bias in estimates of the recombination frequency. Table 7 shows the means and standard deviations of the maximum-likelihood estimates of recombination frequencies for 100 replicate simulations of some pairs of marker loci considered above. Linkage was detected as significant ($P < 0.05$) by both the independence test and the likelihood-ratio test with a frequency ≥90% when the recombination frequency $r \leq 0.3$, except for the least informative pair (L_2, L_5). For $r = 0.5$, the frequency of significant tests was close to 5%. The means of the MLEs of r were close to the corresponding simulated values for $r \leq 0.3$. For $r = 0.5$, the marker estimates were biased downward, due to the selection of the most likely phase. The parental linkage phases at the marker loci were correctly predicted for at least 89% of simulations with cases with $r \leq 0.3$.

LINKAGE ANALYSIS OF EXPERIMENTAL DATA IN AUTOTETRAPLOID POTATO

Some preliminary data from the Scottish Crop Research Institute were used to test this approach, using five SSR marker loci (*STM*0017, *STM*1017, *STM*1051, *STM*1052, and *STM*1102) and six AFLP marker loci (*e*35*m*61–18, *e*35*m*61–21, *e*37*m*39–14, *e*39*m*61–7, *e*46*m*37–12, and *p*46*m*37–12) scored on 77 offspring from a cross between two parental lines: the advanced potato breeding line 12601abl and the cultivar Stirling (Bradshaw *et al.* 1998). Details of scoring the DNA molecular markers are described in Meyer *et al.* (1998) and Milbourne *et al.* (1998). Preliminary analysis of the AFLP markers (Meyer *et al.* 1998), and of the SSR markers in diploid and tetraploid populations (Milbourne *et al.* 1998) suggested that these markers are all on the same linkage group.

Table 8 summarizes the parental phenotypes and the phenotype distribution of the offspring at the marker loci. Of a total of 77 offspring scored at these marker loci, there were 73, 73, 72, and 70 progeny whose phenotypes at the marker loci *STM*1017, *STM*1051, *STM*1052, and *STM*1102, respectively, were unambiguously observed. The phenotypic data were used to predict the parental genotypes using the method of Luo *et al.* (2000). The predicted parental genotypes at the marker loci are also shown in Table 8 together with the corresponding prediction probabilities and the χ^2 values of

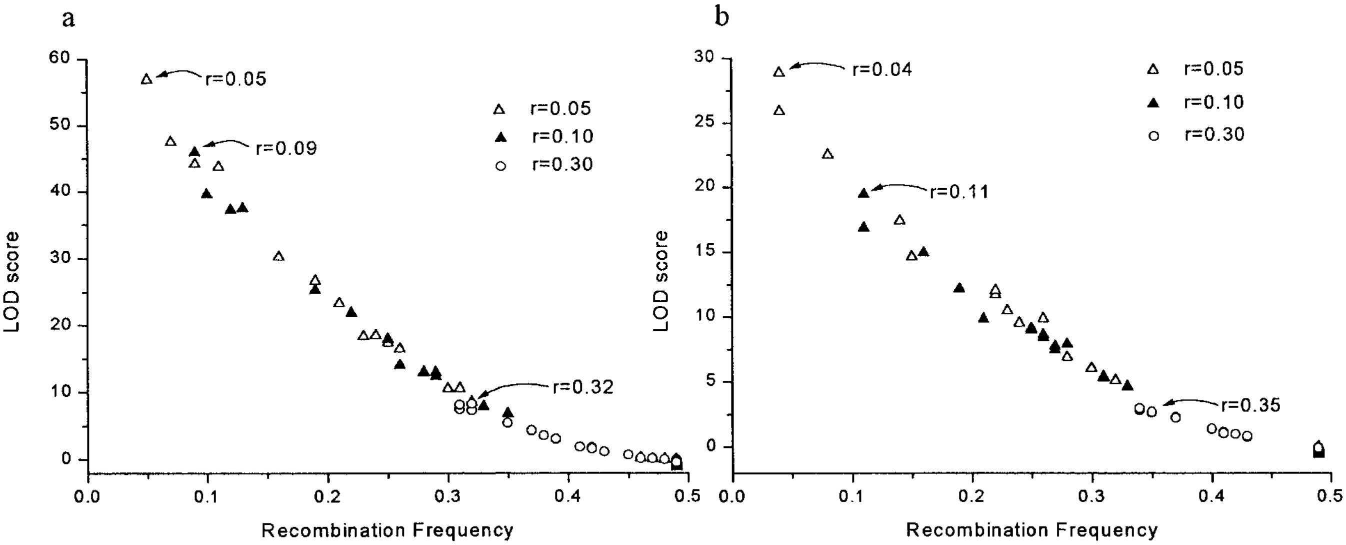


FIGURE 4.—The maximum-likelihood estimates of recombination frequencies and the corresponding LOD scores for all possible linkage phases for (a) loci L_9 and L_{10} and (b) loci L_7 and L_{10}. Arrows indicate the true linkage phases.

the goodness-of-fit test. It was found from the analysis that the number of possible genotype configurations with probability ≥0.1 varies from 1 (at *STM*0017, *STM*1051, and the AFLP marker loci) up to 8 (*STM*1017). For this locus, all that can be deduced is that allele 1 occurs in a simplex condition in parent 1.

The independence tests were performed for all possible pairs of the marker loci and the significance probabilities of the tests are listed as the first rows of the lower diagonal in Table 9, along with the results of a pairwise linkage analysis. The maximum-likelihood estimates of recombination frequency between the pairs of marker loci are listed on the upper diagonal and the LOD scores are given in the second rows of the lower diagonal of Table 9. Both possible genotypes for locus *STM*1052 are shown, as these gave slightly different estimates of the recombination frequencies and LOD scores. The likelihood for genotype *AABO* × *AACO* was always larger than for the other genotype. The use of the alternative parental genotypes at loci *STM*1017 and *STM*1102 did

TABLE 7

Mean and standard deviation of the maximum-likelihood estimate of recombination frequency and the empirical statistical power for detecting the linkage based on 100 simulations

Set	Marker loci	n	r	$\hat{r}$ ± SD	RI	β_1	β_2	ξ
1	L_1 and L_2	100	0.1	0.0969 ± 0.030	0.95	99	99	99
2	L_1 and L_2	200	0.1	0.1051 ± 0.021	0.95	100	100	100
3	L_1 and L_2	200	0.3	0.3027 ± 0.042	0.80	100	95	93
4	L_1 and L_2	200	0.5	0.4197 ± 0.029	0.74	6	4	—
5	L_1 and L_3	100	0.1	0.1012 ± 0.022	1.65	99	99	99
6	L_1 and L_3	200	0.1	0.1002 ± 0.015	1.65	100	100	100
7	L_1 and L_3	200	0.3	0.3030 ± 0.029	1.31	100	100	100
8	L_1 and L_3	200	0.5	0.4473 ± 0.018	1.15	5	4	—
9	L_2 and L_5	100	0.1	0.1053 ± 0.094	0.10	100	75	100
10	L_2 and L_5	200	0.1	0.1060 ± 0.069	0.10	100	97	100
11	L_2 and L_5	200	0.3	0.3057 ± 0.088	0.11	79	55	89
12	L_2 and L_5	200	0.5	0.4062 ± 0.047	0.12	2	0	—
13	L_9 and L_{10}	100	0.1	0.1033 ± 0.029	1.34	100	100	100
14	L_9 and L_{10}	200	0.1	0.1000 ± 0.018	1.34	100	100	100
15	L_9 and L_{10}	200	0.3	0.2948 ± 0.030	1.36	100	100	89
16	L_9 and L_{10}	200	0.5	0.4409 ± 0.025	1.37	3	1	—

RI, the relative information; β_1, the frequency of the significant tests of independence at $\alpha = 0.05$; β_2, the frequency of the likelihood-ratio tests significant with the threshold of 3.84 given the independence test is significant; and ξ, the frequency of correct prediction of the simulated linkage phase given the independence test is significant.

TABLE 8

Phenotypes of five SSR and six AFLP marker loci scored on two parents (P_1, Stirling; P_2, 12601abl) and their progeny and the predicted parental genotypes G_1 and G_2 at these marker loci

Markers	P_1	P_2	n	O_i	f_i	$G_1 \times G_2$	Probability	$\chi^2_{\text{d.f.}}$
M_1 = *STM*1052	1 1 0	1 0 1	72	1 1 1	0.28	*AABO* × *ACOO*	0.11	5.88
				0 1 1	0.04	*AABO* × *AACO*	0.84	3.56
				1 0 0	0.25			
				1 1 0	0.17			
				1 0 1	0.26			
M_2 = *STM*1051	1 1 1 0	0 1 0 1	73	1 1 0 1	0.23	*ABBC* × *BBDO*	0.95	5.90
				1 1 0 0	0.16			
				0 1 0 0	0.04			
				1 1 1 0	0.05			
				0 1 1 0	0.19			
				0 1 0 1	0.11			
				0 1 1 1	0.12			
				1 0 1 1	0.03			
				1 1 1 1	0.05			
M_3 = *STM*0017	0 1	1 0	77	0 1	0.18	*BOOO* × *AOOO*	0.99	4.82
				0 0	0.25			
				1 1	0.35			
				1 0	0.22			
M_4 = *STM*1017	1 1	0 1	73	1 1	0.49	*ABOO* × *BBBO*	0.12	0.01
				0 1	0.51	*ABOO* × *BBBB*	0.12	0.01
						ABBO × *BBBO*	0.12	0.01
						ABBO × *BBBB*	0.12	0.01
						ABBB × *BOOO*	0.12	0.01
						ABBB × *BBOO*	0.12	0.01
						ABBB × *BBBO*	0.12	0.01
						ABBB × *BBBB*	0.12	0.01
M_5 = *STM*1102	0 1 1	1 0 1	70	0 1 1	0.44	*BBCO* × *ACCC*	0.46	2.55
				1 1 1	0.36	*BBCC* × *ACCC*	0.46	2.55
				1 0 1	0.13			
				0 0 1	0.07			
M_6 = *e*35*m*61–18	1	0	77	1	0.53	*AOOO* × *OOOO*	0.99	0.32
				0	0.47			
M_7 = *e*35*m*61–21	1	0	77	1	0.84	*AAOO* × *OOOO*	0.99	0.06
				0	0.16			
M_8 = *e*37*m*39–14	1	0	77	1	0.62	*AOOO* × *OOOO*	0.99	4.69
				0	0.38			
M_9 = *e*39*m*61–7	1	0	77	1	0.51	*AOOO* × *OOOO*	0.99	0.05
				0	0.49			
M_{10} = *e*46*m*37–12	1	0	77	1	0.45	*AOOO* × *OOOO*	0.99	0.01
				0	0.55			
M_{11} = *p*46*m*37–12	1	0	77	1	0.48	*AOOO* × *OOOO*	0.99	0.00
				0	0.52			

O_i and f_i are marker phenotypic classes of the offspring and the corresponding frequencies, respectively.

not affect the estimates of recombination frequencies and LOD scores.

The MLEs of the pairwise recombination frequencies and the LOD scores were used to map the marker loci using JoinMap. The 11 markers were mapped as a linkage group with a length of 48.9 cM (using genotype *AABO* × *AACO* for *STM*1052). The order was the same using the alternative genotype, and the calculated length in this case was 48.7 cM. The allelic linkage phases of the parental genotypes at the marker loci were reconstructed as described. The linkage map and the reconstructed phases are illustrated in Figure 5. For loci *STM*1017 and *STM*1102, the dosage of the alleles that are present for all offspring is uncertain, but the phase of the segregating alleles can be reconstructed. For *STM*0017, there is uncertainty about the phase for parent 2 as this marker is well separated from the other SSR markers that are informative about parent 2, although allele *A* of this marker is unlikely to be linked in coupling to the simplex alleles of the other SSR markers (*A* for *STM*1102, *C* for *STM*1052, and *D* for *STM*1051). The only difference between the inferred phase in Figure 5

TABLE 9

The maximum-likelihood estimates of pairwise recombination frequencies (the upper diagonal) among five SSR and six AFLP marker loci in autotetraploid potato, their corresponding LOD scores (the second rows of the lower diagonal), and the significance level of the independence tests (the first rows of the lower diagonal)

Loci	M_1	$M_{1'}$	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
M_1	—	–	0.028	0.166	0.041	0.003	0.156	0.003	0.111	0.072	0.248	0.012
$M_{1'}$	—	—	0.028	0.201	0.017	0.004	0.084	0.002	0.111	0.029	0.224	0.097
M_2	0.000	0.000	—	0.196	0.105	0.121	0.239	0.110	0.164	0.143	0.285	0.123
	35.08	34.60										
M_3	0.109	0.109	0.014	—	0.274	0.391	0.026	0.150	0.399	0.053	0.104	0.065
	1.75	1.59	2.49									
M_4	0.017	0.017	0.033	0.001	—	0.499	0.274	0.371	0.270	0.278	0.329	0.352
	1.58	1.64	2.94	3.36								
M_5	0.003	0.003	0.017	0.115	0.375	—	0.413	0.208	0.153	0.466	0.398	0.136
	5.50	4.98	4.84	0.90	0.00							
M_6	0.086	0.086	0.114	0.000	0.000	0.407	—	0.074	0.399	0.053	0.104	0.065
	0.89	1.08	1.32	19.15	3.36	0.10						
M_7	0.015	0.015	0.009	0.006	0.351	0.000	0.001	—	0.110	0.178	0.107	0.078
	3.54	3.38	3.28	1.77	0.19	2.42	2.82					
M_8	0.000	0.000	0.000	0.120	0.197	0.005	0.463	0.000	—	0.299	0.175	0.100
	10.77	10.77	7.81	0.07	0.36	2.19	0.07	2.69				
M_9	0.075	0.075	0.036	0.000	0.000	0.721	0.000	0.017	0.211	—	0.079	0.228
	1.37	1.47	2.58	16.07	3.20	0.01	16.07	1.32	0.29			
M_{10}	0.172	0.172	0.063	0.000	0.003	0.371	0.000	0.005	0.071	0.000	—	0.213
	0.49	0.49	0.87	12.02	1.90	0.14	12.02	1.95	0.82	13.76		
M_{11}	0.007	0.007	0.000	0.054	0.407	0.019	0.001	0.001	0.017	0.110	0.080	—
	2.92	2.82	10.14	1.51	0.15	2.10	1.51	2.67	1.26	0.56	0.64	

The order of the marker loci is in accordance with that listed in Table 8. M_1 and $M_{1'}$ represent the two alternative genotypes at locus *STM*1052.

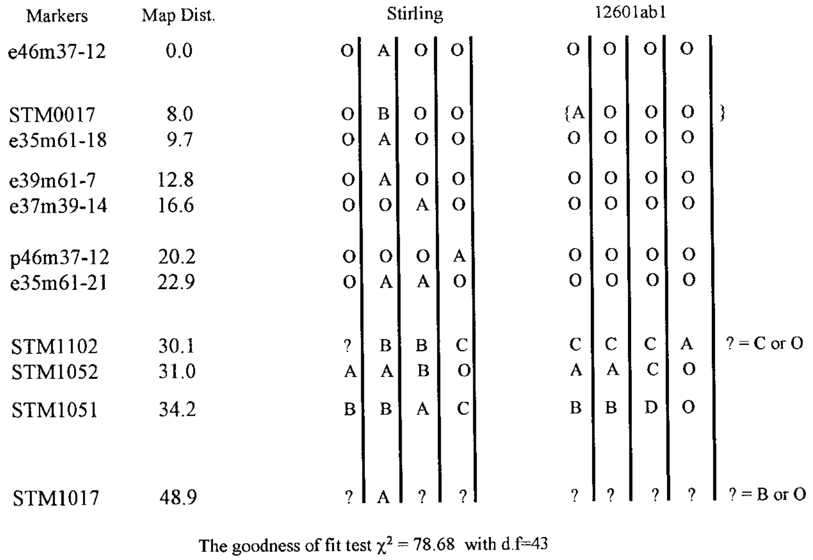
and the most likely phase is for the pair *STM*0017 and *STM*1102, for which the inferred phase has a log-likelihood 1.16 less than the most likely, although both phases correspond to loose linkages (recombination frequency ≈ 0.4, LOD 0.90). The conclusion that the SSR markers form a single linkage group agrees with the analysis of these markers in a diploid cross (Milbourne *et al.* 1998).

DISCUSSION

In this article we have developed the methodology for constructing linkage maps of codominant or dominant genetic markers in autotetraploid species under chromosomal segregation, *i.e.*, the random pairing of four homologous chromosomes to give two bivalents. Our strategy has the following steps:

1. Identify which parental genotype(s) are consistent with the parental and offspring phenotype data.
2. For each pair of loci, calculate Pearson's chi-square statistic for independent segregation, and its significance.
3. Use cluster analysis, based on the significance, to partition the loci into linkage groups. For each linkage group in turn, proceed as follows:
4. For each pair of loci, calculate the recombination frequency and the LOD score for all possible phases. The EM algorithm allows this to be done for any parental genotype configuration.
5. For each pair, identify the phase with the largest likelihood and estimate the difference in log-likelihood d_{ij} between the most likely and second most likely phase.
6. Use the recombination frequencies and LOD scores for the most likely phases to order the loci and calculate distances between them.
7. Reconstruct the linkage phase for the complete linkage group, using pairs in order of decreasing d_{ij}.
8. Check that the inferred linkage phases are the most likely ones for all pairs with a substantial difference in log-likelihood, for example $d_{ij} > 3$.
9. For pairs where the inferred phase is not the most likely phase, compare estimates of the recombination frequencies and LOD scores. Recalculate the linkage map on the basis of the inferred phase if necessary.

In the simulated and experimental data sets, there have been examples of loci for which more than one genotype for the parents is possible. This occurred for three reasons. First, some alleles may be present in all offspring, and so are uninformative, for example, simulated locus L_5. Alleles *A* and *C* are present in parents 1



Markers	Map Dist.	Stirling				12601ab1				
e46m37-12	0.0	O	A	O	O	O	O	O	O	
STM0017	8.0	O	B	O	O	{A	O	O	O	}
e35m61-18	9.7	O	A	O	O	O	O	O	O	
e39m61-7	12.8	O	A	O	O	O	O	O	O	
e37m39-14	16.6	O	O	A	O	O	O	O	O	
p46m37-12	20.2	O	O	O	A	O	O	O	O	
e35m61-21	22.9	O	A	A	O	O	O	O	O	
STM1102	30.1	?	B	B	C	C	C	C	A	? = C or O
STM1052	31.0	A	A	B	O	A	A	C	O	
STM1051	34.2	B	B	A	C	B	B	D	O	
STM1017	48.9	?	A	?	?	?	?	?	?	? = B or O

The goodness of fit test $\chi^2 = 78.68$ with d.f=43

FIGURE 5.—The linkage map and parental linkage phases reconstructed from five SSR and six AFLP markers using a full-sib family from two autotetraploid potato lines. (Stirling and SCRI clone 12601abl). (?) An allele unresolved in the linkage analysis. The phase of marker *STM*0017 in parent 1260lab1 (shown in braces {}) cannot be resolved.

and 2, respectively, and in all offspring; only allele *D* (present in parent 2 only) segregates in the offspring in a 1:1 ratio. The parental genotypes are consistent with either *AAAA* × *CCCD* or *AAAO* × *CCCD* and, as all information about linkage comes from the segregating allele *D*, the choice between these two genotypes has no consequence for the estimation of the map. This will be the situation if all the possible genotypes have the same configuration for the segregating alleles. Second, the parents may have the same phenotypes but different genotypes, *e.g.*, simulated locus L_2 where genotypes *AABC* × *ABCC* and *ABCC* × *AABC* are possible. In this case, comparison of the likelihoods for the two possible genotypes segregating jointly with a linked, informative marker should resolve the issue. For locus L_2, the likelihood of the joint segregation data with loci L_1, L_3, L_4, L_6, and L_7 was consistently higher for the true genotype *AABC* × *ABCC* than for the alternative genotype *ABCC* × *AABC.* Third, the offspring phenotypes may be compatible with more than one possible genotype configuration, with different configurations for the segregating alleles, *e.g.*, STM1052. In this case, the best approach is to calculate and compare the maps using each genotype. For the experimental data used here, the differences in the maps were negligible, but this may not always be so.

The examination of the information of different configurations shows that, as expected, there are many configurations of codominant markers that are more informative than the simplex coupling configuration, which is the most informative configuration for a dominant marker, as demonstrated in HACKETT *et al.* (1998). Markers with many different alleles are most informative, and markers with multiple doses of alleles or alleles shared by both parents are less informative in general, but linkage phase also contributes, and so it is difficult to reject any locus configuration as uninformative for mapping purposes.

The reconstruction of the parental genotypes involves a test for double reduction. LUO *et al.* (2000) showed that the power of this test was high for detecting double reduction, but no significant double reduction was found in the experimental data. Little work has been done on the theory for predicting the joint segregation probabilities under a two-loci tetrasomic inheritance model when double reduction occurs, and we have not attempted to include it in the linkage analysis at present. However, double reduction is known to occur in potato (BRADSHAW and MACKAY 1994). It has also been observed that in potato, while bivalents predominate, low frequencies of quadrivalents, trivalents, and univalents occur (SWAMINATHAN and HOWARD 1953). In autotetraploid alfalfa, in contrast, BINGHAM and MCCOY (1988) found that most cells have the full complement of 16 bivalents at metaphase I. We hope to explore these complications in a future publication. In the meantime it is worth exploring the use of the current simple model on as wide a range of real data as possible.

Inference of linkage phase is a complicated issue in linkage analysis for diploids and even more so for polyploids, particularly when multiple loci have to be considered simultaneously. In this study, a likelihood-based approach was proposed to search over all possible linkage phase configurations of any given pair of tetraploid parental genotypes at two loci for the most likely one. For closely linked and/or informative pairs of loci, the difference between the most likely and the second most likely phase is clear-cut, and then the actual phase was predicted adequately. However, several phases may be nearly equally likely when the loci are loosely linked or the genotypic pair is less informative. In the cases examined here, phases with similar likelihoods had similar inferred recombination frequencies. Because of this, it is reasonable to calculate the linkage map using the recombination frequencies and LOD scores for the most likely phases for each pair, reconstruct the phase for the whole group, and then compare the estimates of the recombination frequencies at the inferred and most likely phases where these differ. For the simulated data, the difference in the estimates of the recombination frequency was always <0.01. We did not find any case where a difference in phase between the inferred and most likely one caused a nonnegligible difference in the estimate of the recombination frequency, but the possibility of this should be borne in mind.

This analysis has reconstructed the map on the basis of pairwise analyses. *A* least-squares method, implemented in the JoinMap software, was used to calculate multipoint map distances. A practical strategy is suggested for constructing the phase for the entire linkage group from the estimated pairwise phases and for checking its consistency. This approach reconstructed the complete phases correctly for the three linkage groups of our simulation study and gave a consistent phase for the experimental data. In theory, our approach could be improved by the use of a multilocus linkage analysis and phase analysis. There have been several approaches to multilocus linkage analysis in diploids. Prominent among them is the hidden Markov chain model proposed by Lander and Green (1987). The multilocus approach takes into consideration the cosegregation of genes at several linked loci simultaneously, and problems such as missing marker data and incomplete information of some markers (for example, dominant markers) can be appropriately addressed in the analysis. The basic principle of the multilocus linkage analysis in diploids would be extendable to tetraploids, but innovative theoretical efforts would have to be invested to model a more complicated stochastic process of multilocus crossovers under tetrasomic inheritance, and more efficient numerical algorithms have to be developed to analyze the model. In particular, methods have to be developed to handle the large number of possible linkage phases, which for a tetraploid genotype has a maximum of $4!^{m-1}$ for m linked loci, or $4!^{2(m-1)}$ phases for the two parents. A possible way to tackle the problem might be the use of the Markov property of recombinant events over the linked marker loci as demonstrated for diploids in Jiang and Zeng (1997). The Markov model allows the division of all marker loci under question into groups flanked by fully informative markers, thus reducing the scale of the modeling problem. In practice, the scarcity of fully informative markers (*i.e.*, with eight distinct alleles) in tetraploid species may be a difficulty.

A direct utility of the marker linkage maps is to map QTL. Doerge and Craig (2000) recently proposed a model selection strategy for quantitative trait locus analysis in polyploids. Though their method was suitable only for a single-marker QTL linkage test, the study highlighted aspects of difficulties and tools necessary to investigate QTL mapping in polyploids. As they have recognized, QTL analysis under a setting of multiple-marker loci will present entirely new challenges. The methodologies developed in the present article open another window for viewing and tackling the complexities of polyploid linkage analysis with quantitative trait loci.

We thank two anonymous reviewers and Dr. Z-B. Zeng for the comments and criticisms that have been very helpful in improving the manuscript. We are grateful for useful discussions with Dr. R. C. Meyer. This research was financially supported by a research grant from the UK Biotechnology and Biological Sciences Research Council. Z.W.L. was also supported by China's "973" program, the National Science Foundation, the QiuShi Foundation, and the Changjiang Scholarship; the other authors were supported by the Scottish Executive Rural Affairs Department.

LITERATURE CITED

Agresti, A., 1990 *Categorical Data Analysis*. Wiley, New York.

Al-Janabi, S. M., R. J. Honeycutt, M. McClelland and B. W. S. Sobral, 1993 A genetic linkage map of *Saccharum spontaneum* L. "SES 208." Genetics **134:** 1249–1260.

Bailey, N. T. J., 1961 *Introduction to The Mathematical Theory of Genetic Linkage*. Clarendon Press, Oxford.

Bingham, E. T., and T. J. McCoy, 1988 Cytology and cytogenetics of alfalfa, pp. 737–776 in *Alfalfa and Alfalfa Improvement*, edited by A. A. Hanson. Agronomy Monographs 29, ASA, CSSA, and SSSA, Madison, WI.

Bonierbale, M. W., R. L. Plaisted and S. D. Tanksley, 1988 RFLP maps based on a common set of clones reveal modes of chromosomal evolution in potato and tomato. Genetics **120:** 1095–1103.

Bradshaw, J. E., and T. J. Mackay, 1994 *Potato Genetics*. CAB International, Wallingford, UK.

Bradshaw, J. E., C. A. Hackett, R. C. Meyer, D. Milbourne, J. W. McNicol *et al.*, 1998 Identification of AFLP and SSR markers associated with quantitative resistance to *Globodera pallida* (Stone) in tetraploid potato (*Solanum tuberosum* subsp. *tuberosum*) with a view to marker-assisted selection. Theor. Appl. Genet. **97:** 202–210.

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.

Callen, D. F., A. D. Thompson, Y. Shen, H. A. Phillips, R. I. Richards *et al.*, 1993 Incidence and origin of "null" alleles in the $(AC)_n$ microsatellite markers. Am. J. Hum. Genet. **52:** 922–927.

Cressie, N., and T. R. C. Read, 1984 Multinomial goodness-of-fit tests. J. R. Stat. Soc. Ser. B **46:** 440–464.

Crosby, J. L., 1973 *Computer Simulation in Genetics*. Wiley, New York.

Da Silva, J. A. G., 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 (with discussion). J. R. 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:** 121–124.

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. Lond. Ser. B **23:** 55–87.

Gebhardt, C., E. Ritter, T. Debener, U. Schachtschabel, B. Walkemeier *et al.*, 1989 RFLP-analysis and linkage mapping in *Solanum tuberosum*. Theor. Appl. Genet. **78:** 65–75.

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

Jiang, C., and Z-B. Zeng, 1997 Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica **101:** 47–58.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

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

Luo, Z. W., R. Thompson and J. A. Woolliams, 1997 A population genetics model of marker-assisted selection. Genetics **146:** 1173–1183.

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.

Maliepaard, C., J. Jansen and J. W. Van Ooijen, 1997 Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications. Genet. Res. **70:** 237–250.

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. McNicol *et al.*, 1998 Linkage analysis in tetraploid potato and associations of markers with quantitative resistance to late blight (*Phytophthora infestans*). Mol. Gen. Genet. **259:** 150–160.

Milbourne, D., R. C. Meyer, A. J. Collins, L. D. Ramsay, C. Gebhardt *et al.*, 1998 Isolation, characterisation and mapping simple sequence repeat loci in potato. Mol. Gen. Genet. **259:** 233–245.

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

Stam, P., and J. W. Van Ooijen, 1995 *JoinMap Version 2.0: Software for the Calculation of Genetic Linkage Maps.* CPRO-DLO, Wageningen, The Netherlands.

Swaminathan, M. S., and H. W. Howard, 1953 The cytology and genetics of the potato (*Solanum tuberosum*) and related species. Bibliogr. Genet. **16:** 1–192.

Terwilliger, J. D., Y. L. Ding and J. Ott, 1992 On the relative importance of marker heterozygosity and intermarker distance in gene mapping. Genomics **13:** 951–965.

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: Z-B. Zeng

APPENDIX: DERIVATION OF THE INFORMATION MEASURE

By definition,

$$I(r) = -E\frac{d}{dr}\left[\sum_{i=1}^{k}\frac{n_i}{f_i}\frac{d}{dr}(f_i)\right] = E\sum_{i=1}^{k}n_i\left[\frac{1}{f_i}\frac{d^2f_i}{dr^2} - \frac{1}{f_i^2}\left(\frac{df_i}{dr}\right)^2\right].$$

The expectation of the first item on the right-hand side of the above equation is zero, and then

$$\begin{aligned}
E\sum_{i=1}^{k}n_i\frac{1}{f_i^2}\left(\frac{df_i}{dr}\right)^2 &= E\sum_{i=1}^{k}\frac{n_i}{f_i^2}\left(\sum_{j=0}^{4}\frac{dz_{ij}}{dr}\right)^2 = E\sum_{i=1}^{k}\frac{n_i}{f_i^2}\left(\sum_{j=0}^{4}z_{ij}\frac{d(\log z_{ij})}{dr}\right)^2\\
&= E\sum_{i=1}^{k}\frac{n_i}{f_i^2}\left(\sum_{j=0}^{4}\frac{z_{ij}(j-4r)}{r(1-r)}\right)^2\\
&= E\left[\frac{1}{r^2(1-r)^2}\sum_{i=1}^{k}\frac{n_i}{f^2}\left(\sum_{j=0}^{4}jz_{ij} - 4r\sum_{j=0}^{4}z_{ij}\right)^2\right]\\
&= E\left[\frac{1}{r^2(1-r)^2}\sum_{i=1}^{k}\frac{n_i}{f_i^2}\left(\sum_{j=0}^{4}jz_{ij} - 4rf_i\right)^2\right].
\end{aligned}$$

Substituting $E(n_i) = nf_i$, we derive the information measure as

$$\begin{aligned}
I(r) &= \frac{1}{r^2(1-r)^2}\sum_{i=1}^{k}\frac{n}{f_i}\left(\sum_{j=0}^{4}jz_{ij} - 4rf_i\right)^2\\
&= \frac{n}{r^2(1-r)^2}\sum_{i=1}^{k}\frac{1}{f_i}\left[\left(\sum_{j=0}^{4}jz_{ij}\right)^2 - 8rf_i\sum_{j=0}^{4}jz_{ij} + 16r^2f_i^2\right]\\
&= \frac{n}{r^2(1-r)^2}\left[\sum_{i=1}^{k}\frac{1}{f_i}\left(\sum_{j=0}^{4}jz_{ij}\right)^2 - 16r^2\right],
\end{aligned}$$

noting that $\Sigma_{i=1}^{k}\,\Sigma_{j=0}^{4}jz_{ij} = \Sigma_{j=0}^{4}j\binom{4}{j}r^j(1-r)^{4-j} = 4r$ and $\Sigma_{i=1}^{k}f_i = 1$.