Theor Appl Genet (2000) 100:1067–1073

Z.W. Luo · C.A. Hackett · J.E. Bradshaw
J.W. McNicol · D. Milbourne

Predicting parental genotypes and gene segregation for tetrasomic inheritance

Received: 7 June 1999 / Accepted: 28 September 1999

Abstract Recent genome mapping projects in tetraploid plant species require a method for analysing the segregation patterns of molecular marker loci in these species. The present study presents a theoretical model and a statistical analysis for predicting the genotypes of a pair of tetraploid parents at a codominant (for example, RFLPs, microsatellites) or dominant (for example, AFLPs, RAPDs) molecular marker locus based on their and their progeny's phenotypes scored at that locus (gel-band patterns). The theory allows for null alleles and for any degree of double-reduction to be modelled. A simulation study was performed to investigate the properties of the theoretical model. This showed that in many circumstances both the parental genotypes can be correctly identified with a probability of nearly 1, even when the molecular data were complicated by null alleles or double-reduction. Configurations where the parental genotype cannot be identified are discussed. The power to detect double-reduction varies considerably, depending on the proportion of identical alleles carried and shared by the parents, and the number of null alleles. Incorrect deductions of the occurrence of double-reduction were rare. The method was applied to data on a microsatellite locus segregating in the parents and 74 offspring of a tetraploid potato cross. Twentyfour parental configurations were consistent with the parental gel pattern, but only one of these was compatible with all the phenotypic data on the offspring. The feasibility for extending the present model to predict segregation of several linked loci, and particularly the linkage phase, is briefly discussed.

Communicated by J.W. Snape

Z.W. Luo · C.A. Hackett (✉) · J.W. McNicol
Biomathematics and Statistics Scotland, SCRI, Invergowrie,
Dundee, DD2 5DA, Scotland, UK
e-mail: chacke@scri.sari.ac.uk

J.E. Bradshaw · D. Milbourne
Scottish Crop Research Institute, Invergowrie, Dundee,
DD2 5DA, Scotland, UK

Present address: Z.W. Luo,
Institute of Genetics, Fudan University,
Shanghai 200433, China

Key words Autotetraploid · Potato · Codominant marker · Genotype prediction · Mapping

Introduction

Understanding the genetic mechanisms of polyploidy has long been considered an important aspect of plant genetic research because of its implications in the evolutionary biology of plants and for crop improvement. Polyploidy has played an important role in the establishment of reproductive isolation in many plant species (Lewis 1980). Grant (1971) has estimated that 47% of angiosperms are polyploids. In particular, a large proportion of cultivated crops are allopolyploids which display disomic inheritance, perhaps as many as 50% (Briggs and Knowles 1967). Prominent among them are wheat, cotton, tobacco and many of the forage grasses. Autopolyploids which display polysomic inheritance are less common, but include some important crops such as potato (tetraploid), alfalfa (tetraploid), sugarcane (octoploid) and strawberry (octoploid), none of which are grown for their seed. This reflects the fact that autopolyploids tend to have larger plant parts but reduced fertility. Genetic studies of autopolyploid species are much less advanced than for diploid and allopolyploid species. This is partially due to the complexities of their polysomic inheritance: a substantially wider range of genetic segregation types occurs at individual loci and double-reduction can occur (Mather 1936).

Rapid developments in modern molecular genetics and computer technology have made both theoretical and experimental analyses of polysomic inheritance much more feasible than ever before (De Winton and Haldane 1931; Fisher 1947). DNA genetic markers have recently been used as a fundamental tool to construct genetic linkage maps in polyploid species (Al-Janabi et al. 1993; Da Silva et al. 1993; Hackett et al. 1998), to search for quantitative trait loci (QTLs) controlling 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). A strong assumption made in these studies was that genotypes at the marker loci are expected to be feasibly and accurately predicted from their phenotype. This can be true only in very limited cases (see the following discussion). There are several reasons for a simple one-to-one relationship not to exist between marker genotype and phenotype which, currently, is generally scored as a gel-band pattern. Firstly, where polysomic inheritance is concerned, a multiple dosage of alleles cannot be straightforwardly distinguished from the gel-band pattern. Secondly, some alleles may not be revealed as the presence of a corresponding gel-band because of a failure of the PCR primers to anneal to the relevant DNA templates properly (i.e. so called “null” alleles). This generally happens due to a high mutation within primer sequences (Callen et al. 1993). Thirdly, if the marker loci are far from the centromere, alleles at these loci may show a distorted segregation pattern due to double reduction, the phenomenon that sister chromatids can end in the same gamete as a result of homologous chromosomes forming a multivalent and then crossing-over occurring between the locus and spindle attachment (Bailey 1961). A further potential source of confusion in this type of analysis is the possible occurrence of multi-locus markers, with segregating bands arising from two or more genetically resolvable, independent loci. This phenomenon may be a consequence of the association of the marker (or its flanking regions) with multigene families (in the case of genic sequences), or repetitive, mobile or duplicated elements in the genome.

The aim of the present study is to develop the theory and methods for predicting the genotypes of a pair of tetraploid individuals at a genetic marker locus from their phenotypes and the phenotypes of their offspring. This is a prerequisite for linkage analysis of the genetic marker loci since information about parental genotypes is essential for distinguishing recombinant and parental genotypic classes. The theoretical model takes account of all aspects of the complexities discussed above. Properties of the model and its predictive ability are investigated using simulated data, and finally the methods are demonstrated using experimental data from a tetraploid potato study.

Theory

Model and notation

We will consider the segregation of a single codominant marker locus of an autotetraploid species. Let G_1 and G_2 be two genotypes at the locus for two parental individuals respectively. G_i (i = 1,2) can be expressed as a vector with a length of 4. Because any two tetraploid individuals may have at most eight unique alleles, each element of G_i may take any integer between 0 and 8 where 0 represents the null-allele. Let P_1 and P_2 be the phenotypes of the two individuals, i.e. their pattern of gel bands at the marker locus. P_i (i = 1,2) can be denoted by a vector with length 8, each of whose elements may take a value of 1 indicating the presence of a band at the corresponding gel position, or 0 indicating absence of a band.

Table 1 summarizes the relationship between genotype and phenotype at the marker locus, taking into account the possibility of null-alleles and multiple dosages of identical alleles. It can be seen from Table 1 that there may be 4, 6, 4 or 1 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.

Let α be the coefficient of double-reduction, which is defined as the probability of two sister chromatids occurring in the same gamete during the tetrasomic meiosis. This probability depends on the location of the locus relative to the centromere. The coefficient takes a value varying from 0 (i.e. double-reduction is absent) to its maximum value of 1/6 (Mather 1936; Fisher and Mather 1943). The distribution of gametes produced from an individual with a genotype of $A_1A_2A_3A_4$ is a function of the coefficient of double-reduction, and can be summarized by expressing the frequency of gamete $A_i\mathrm{A}_j$ as $P(A_iA_j) = (1–\alpha) / 6$ ($i \neq j$) or $P(A_iA_i) = \alpha/4$ (Bailey, 1961, p106–108).

Calculation of parental genotypic distribution

Let $O = (o_1, \ldots, o_n)$ be the marker phenotypes of the n offspring from a cross between the two parents. We wish to calculate the probabilities of the possible parental genotypes given the parental and offspring phenotypes and given the value of the coefficient of double-reduction, i.e. $Pr(G_1, G_2|P_1, P_2, O, \alpha)$.

By Bayes’ theorem,

$$Pr(G_1,G_2|P_1,P_2,O,\alpha) = \frac{Pr(O|G_1,G_2,P_1,P_2,\alpha)Pr(G_1,G_2,P_1,P_2,\alpha)}{\sum_u \sum_v Pr(O|G_1=u,G_2=v,P_1,P_2,\alpha)Pr(G_1=u,G_2=v,P_1,P_2,\alpha)}.$$

Now $Pr(O|G_1, G_2, P_1, P_2, \alpha) = Pr(O|G_1, G_2, \alpha)$ since parental phenotypes give no extra information when their genotypes are known, and $Pr(\mathrm{G}_1, G_2, P_1, P_2, \alpha) = Pr(\mathrm{G}_1, G_2|P_1, \mathrm{P}_2) \times Pr(P_1, P_2, \alpha)$. Thus the above equation can be simplified to

$$Pr(G_1,G_2|P_1,P_2,O,\alpha) = \frac{Pr(O|G_1,G_2,\alpha)Pr(G_1,G_2|P_1,P_2)}{\sum_u \sum_v Pr(O|G_1=u,G_2=v,\alpha)Pr(G_1=u,G_2=v|P_1,P_2)}. \quad (1)$$

Calculation of $Pr(O|G_1, G_2, \alpha)$ is tedious because of the need to calculate the expected distribution of genotypes, and in turn phenotypes, of the offspring given their parental genotypes and the coefficient of double-reduction. Moreover, there may be more than one genotypic configuration consistent with the phenotypes of the two parents. Instead of working out these progeny genotypic distributions one by one algebraically, as in Fortini and Barakat (1980), a computer program was designed to calculate this genotypic distribution for any pair of parental genotypes. To achieve this, evaluation of equation (1) can be decomposed into the following steps:

1) For any observed phenotypes of the two parents, all possible genotypic configurations corresponding to these phenotypes can be determined from Table 1. For a given G_1, G_2 and α, the gamete distributions for each of the parental genotypes are worked out and a random union among these gamete pools is then assumed to produce the progeny genotypic distribution.
2) Given the progeny genotypic distribution, the corresponding phenotypic distribution is thus easily derived from relating all possible genotypes to their phenotypes according to Table 1. Let K be the set of possible phenotype classes with k members and let f_i, $i = 1, 2, \ldots, k$ represent the probability of the ith phenotypic class.
3) Let M be the set of the observed phenotype classes in the offspring. If some members of M are not in the set K of theoretically possible phenotypes, then the given (G_1, G_2, α) is rejected. Otherwise we can calculate the likelihood of the observed offspring phenotypes by assuming that they are a random sample from a multinomial distribution with probabilities f_i, $i = 1$, …, k and sample size $n = \sum_i^k n_i$, where n_i is the observed number of offspring in the ith phenotype class, i.e.

$$L(G_1, G_2, \alpha|O) =$$

$$Pr(O|G_1, G_2, \alpha) = \begin{bmatrix} \binom{n}{n_1, n_2, \ldots, n_k} \; f_1^{n_1} f_2^{n_2} \ldots f_k^{n_k} & M \subset K \\ 0 & M \not\subset K \end{bmatrix}. \tag{2}$$

The most likely pair of parental genotypes is determined by examining equation (1) over all possible genotypic configurations (G_1, G_2) and for a range of values of α between 0 and 1/6.

Detecting and estimating the coefficient of double-reduction

The model discussed above provides a direct test for the presence of double-reduction and an estimate of the coefficient of double-reduction. In fact, for a given pair of parental genotypes, the likelihood function (2) can be readily calculated at any fixed grid of values of α between 0 and 1/6. If (G_1, G_2) = (r*, s*) are the most-likely parental genotypes determined as described above, the maximum-likelihood estimate (MLE) of the coefficient of double-reduction is determined by a value of α, say α*, which maximises the likelihood profile $L[(G_1, G_2) = (r^*, s^*), \alpha|O]$.

Moreover, when the sample size n is sufficiently large, the test statistic

$$\chi^2_{(1)} = 2\{log\ L[(G_1, G_2) =$$

$$(r^*, s^*), \alpha =^* |O] - log\ L[(G_1, G_2) = (r^*, s^*), \alpha = 0|O]\} \tag{3}$$

asymptotically follows a chi-square distribution with 1 degree of freedom and therefore provides a simple significance test for double-reduction.

Simulation study

To validate the theoretical analyses represented above and to investigate their properties, a series of computer programs (in Fortran-90) were written to mimic the tetrasomic inheritance of alleles at a single marker locus and to perform numerical analyses of the simulated data according to equations (1), (2) and (3) above. The programs simulate meiosis in a tetraploid individual with any genotype at the locus (with or without double-reduction), a random sampling of gametes from the meiosis, a random union of gametes randomly sampled from the gamete pool, and the generation of the phenotype from any given individual genotype. An example of a simulated data set is given in Table 2.

Various sample sizes and degrees of double-reduction can be considered in the simulation study. There are a large number of possible pairs of parental genotypes as there are up to nine alleles (including the null-allele) allowed to segregate at the marker locus. For the purpose of demonstration, only a fraction of all these possibilities was considered in the present simulation study. In addition to selecting various parental genotype configurations, the progeny size and the coefficient of double-reduction were also varied. Each simulation configuration was run 100 times. Summarized from the simulation data analyses were: (1) frequencies of the correct prediction of both parental genotypes (denoted by ξ_{12}), the correct prediction of only the first parent but not the second (ξ_1) or the correct prediction of only the second parent but not the first (ξ_2); (2) mean values of the maximum-likelihood estimates of the coefficient of double-reduction (α*) and the corresponding standard deviations over the repeated simulation trials, and (3) the empirical power of

Table 1 Relationship between marker phenotypes and genotypes at a single locus. Different integers represent different alleles, and zero denotes a null allele

Band patterns	Phenotype	Corresponding genotypes
One band	-	(1000), (1100), (1110), (1111)
Two bands	- -	(1200), (1120), (1220), (1122), (1222), (1112)
Three bands	- - -	(1230), (1233), (1223), (1123)
Four bands	- - - -	(1234)

Table 2 Hypothetical band patterns and the corresponding phenotypic records of the bands of two parental tetraploid individuals and ten of their progeny. The data demonstrated here were generated from crossing two individuals with genotypes 1230 and 4567. The coefficient of double-reduction α was equal to 0.1

Individuals	Band no.								Phenotypic record							
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
P_1	–	–	–						1	1	1	0	0	0	0	0
P_2				–	–	–	–		0	0	0	1	1	1	1	0
o_1	–				–		–		1	0	0	0	1	0	1	0
o_2		–					–		0	1	0	0	0	0	1	0
o_3	–	–		–			–		1	1	0	1	0	0	1	0
o_4	–	–		–		–			1	1	0	1	0	1	0	0
o_5			–		–		–		0	0	1	0	1	0	1	0
o_6	–		–	–		–			1	0	1	1	0	1	0	0
o_7		–		–	–				0	1	0	1	1	0	0	0
o_8	–		–		–				1	0	1	0	1	0	0	0
o_9	–	–			–		–		1	1	0	0	1	0	1	0
o_{10}		–	–				–		0	1	1	0	0	0	1	0

Table 3 Analysis of 100 replicate simulations of some parental configurations with four observable alleles in each parent. G_1 and G_2 denote the actual parental genotypes, n is the progeny sample size and α and α^* are respectively the actual value and the mean (standard deviations in parentheses) of the MLEs of the coefficients of double-reduction. β is the empirical power of the statistical test for double-reduction

Configuration	G_1	G_2	n	α	α^*	β
1	1234	5678	50	0.10	0.097 (0.063)	1.00
2	1234	4567	50	0.10	0.098 (0.033)	1.00
3	1234	3456	50	0.10	0.103 (0.041)	0.97
4	1234	2345	50	0.10	0.100 (0.049)	0.85
5	1234	1234	50	0.10	0.095 (0.052)	0.30
6	1234	1234	100	0.10	0.102 (0.046)	0.60

the double-reduction test which was calculated as the frequency of significant tests over simulation replicates. These are illustrated in Tables 3, 4 and 5.

Table 3 shows configurations where the parents have four different alleles and share a varying number of alleles. In these cases the parental genotypes are obtained immediately from the banding patterns. Our analytical approach also reconstructs the correct parental genotype, but its importance here is in the estimation of the coefficient of double-reduction. This coefficient is estimated accurately in each case, but the power of detection dropped from 100% to 30% as the number of shared alleles increased from zero to four. Configuration 6 is identical to configuration 5 but with a sample size of 100: increasing the sample size increased the power to detect double-reduction from 30% to 60%. A significant test for double-reduction when none was simulated occurred for 1 of 100 simulations with four shared alleles, and did not occur for any simulations of the other configurations in Table 3.

In Table 4 and configurations 1–5, both parents carried multiple doses of identical alleles and one of the parents contained varying numbers of null-alleles. The parental genotypes were predicted correctly with a frequency of nearly 100%. Double-reduction was accurately estimated and detected in about 90% of the cases. These were achieved with a progeny size of 100. Configurations 6–9 were used to investigate the effect of the null-allele on the efficiency of the genotype prediction and the power of inferring double-reduction. It is clear from Table 4 that the parent genotypes have been correctly identified in all these simulations and the coefficient of double-reduction was usually well estimated from the maximum-likelihood analysis. However the power for detecting double-reduction was low (< 40%) and decreased as the number of the null-alleles increased.

The last two configurations (10 and 11) of Table 4 are representative of a set of configurations with intrinsic difficulties: although the parents have different genotypes (1230) and (1233) they have the same phenotype, 123. In this case 80/100 simulations with no double-reduction found that parental configurations (1230, 1233) and (1233,1230) were jointly most probable. To distinguish between these would require information on the joint segregation of this locus with one or more linked, highly informative loci.

When the parental phenotypic configuration has only two alleles observable as bands, determining the parental genotype may not be possible. For example, in the absence of double reduction, parental configurations (1111, 1122), (1111, 1220), (1110, 1122) and (1110, 1220) will all give an expected offspring phenotype distribution with two classes: 1/6 with phenotype band 1 only and 5/6 with phenotype bands 1 and 2. These four configurations cannot be distinguished. The only reliable conclusion for this locus is that the allele producing band 2 is present in the second parent in a duplex state, and can be regarded as a dominant marker for linkage analysis. Other parental configurations will give expected offspring phenotypic distributions with the same two classes but different frequencies, but no two-class offspring distribution is associated with a unique parental distribution.

Some parental configurations with two alleles give rise to offspring phenotypic distributions with three or four classes e.g. parents (1122, 1122) produce three classes of offspring (1/36 with band 1 only, 1/36 with band 2 only and 34/36 with bands 1 and 2 in the absence of double-reduction). Assuming no double-reduction, there are ten possible three-class distributions, of which six are associated with a unique parental configuration, two are

Table 4 Analysis of 100 replicate simulations of some parental configurations with three observable alleles. G_1 and G_2 denote the actual parental genotypes, n is the progeny sample size and α and α^* are respectively the actual value and the mean (standard deviations in parentheses) of the MLEs of the coefficients of double-reduction.ξ_{12}, ξ_1 and ξ_2 represent the observed proportions of correct predictions of both G_1 and G_2, G_1 but not G_2, and G_2 but not G_1 respectively. β is the empirical power of the statistical test for double reduction

Configuration	G_1	G_2	n	α	α^*	ξ_{12}	ξ_1	ξ_2	β
1	1220	1333	100	0.00	0.000 (0.000)	0.99	0.00	0.01	0.00
2	1220	1333	100	0.05	0.057 (0.037)	0.96	0.00	0.04	0.89
3	1220	1333	100	0.10	0.099 (0.036)	0.98	0.00	0.02	0.99
4	1200	1333	100	0.00	0.000 (0.000)	1.00	0.00	0.00	0.00
5	1200	1333	100	0.10	0.096 (0.047)	0.99	0.00	0.01	0.91
6	1100	2233	100	0.10	0.087 (0.052)	1.00	0.00	0.00	0.32
7	1000	2233	100	0.10	0.103 (0.049)	1.00	0.00	0.00	0.34
8	1100	2300	100	0.10	0.085 (0.058)	1.00	0.00	0.00	0.20
9	1000	2300	100	0.10	0.093 (0.061)	1.00	0.00	0.00	0.10
10	1230	1233	100	0.00	0.023 (0.042)	0.80	0.00	0.20	0.02
11	1230	1233	100	0.10	0.106 (0.056)	0.84	0.00	0.16	0.64

Table 5 Analysis of 100 replicate simulations of some parental configurations with two observable alleles. G_1 and G_2 denote the actual parental genotypes, n is the progeny sample size, k is the number of offspring phenotype classes in the absence of double reduction, and α and α^* are respectively the actual value and the mean (standard deviations in parentheses) of the MLEs of the coefficients of double reduction. ξ_{12}, ξ_1 and ξ_2 represent the observed proportions of correct predictions of both G_1 and G_2, G_1 but not G_2, and G_2 but not G_1, respectively. β is the empirical power of the statistical test for double-reduction

Configuration	G_1	G_2	n	k	α	α^*	ξ_{12}	ξ_1	ξ_2	β
1	1100	2200	100	4	0.00	0.047 (0.051)	0.99	0.01	0.00	0.01
2	1100	2200	500	4	0.00	0.011 (0.020)	1.00	0.00	0.00	0.03
3	1100	2200	100	4	0.10	0.100 (0.056)	1.00	0.00	0.00	0.24
4	1100	2200	500	4	0.10	0.096 (0.035)	1.00	0.00	0.00	0.75
5	1100	1200	100	4	0.00	0.067 (0.066)	0.63	0.11	0.09	0.10
6	1100	1200	500	4	0.00	0.033 (0.042)	0.93	0.05	0.00	0.03
7	1100	1200	100	4	0.10	0.110 (0.058)	0.75	0.05	0.01	0.20
8	1100	1200	500	4	0.10	0.098 (0.041)	0.97	0.00	0.00	0.60
9	1122	1122	100	3	0.00	0.076 (0.056)	0.46	0.08	0.00	0.01
10	1122	1122	500	3	0.00	0.056 (0.064)	0.72	0.00	0.01	0.03
11	1122	1122	100	3	0.10	0.069 (0.059)	0.57	0.18	0.04	0.00
12	1122	1122	500	3	0.10	0.097 (0.045)	0.98	0.02	0.00	0.53
13	1100	1220	100	2	0.10	0.084 (0.059)	0.47	0.07	0.14	0.15
14	1100	1220	500	2	0.10	0.105 (0.038)	0.84	0.11	0.01	0.88
15	1100	1120	100	2	0.10	0.098 (0.061)	0.00	0.10	0.01	0.07
16	1100	1120	500	2	0.10	0.096 (0.052)	0.14	0.32	0.02	0.94

Table 6 Segregation analysis at an SSR marker locus in tetraploid potato. The left panel of the Table shows the observed band pattern for parents P_1 (12601ab1) and P_2 (Stirling) and the offspring classes O_i (i = 1, ...,8) together with their observed frequencies (Observed) and those expected for the most probable configuration (1334, 2244) (Expected). The right panel of the Table lists predicted parental genotypes and the corresponding probabilities and values of the χ^2 goodness–of–fit test. The coefficient of double-reduction α is fixed at its maximum-likelihood estimate of 0.0

Class	Band pattern				Observed	Expected	Parental genotypes		Probability	$\chi^2_{df=7}$
P_1	1	0	1	1	–	–	1340	2400	0.000000000000	269.1
P_2	0	1	0	1	–	–	1340	2440	0.000000000000	134.0
O_1	0	1	1	1	26	28.8	1340	2240	0.000000000000	122.9
O_2	1	1	1	1	22	16.4	1340	2244	0.000000000000	58.1
O_3	1	1	0	1	9	10.3	1334	2400	0.000000000000	113.0
O_4	1	0	1	1	6	4.1	1334	2440	0.000000000204	49.9
O_5	0	0	1	1	5	6.2	1334	2240	0.000003158273	35.2
O_6	1	1	1	0	3	4.1	1334	2244	0.999996841522	6.2
O_7	0	1	1	0	3	2.1				
O_8	1	0	0	1	0	2.1				

unique if parents are permuted and two can be derived from two different parental configurations. There are also six four-class configurations, each associated with a single parental configuration. If double-reduction occurs, more (but not all) of the offspring distributions will be associated with a unique parental configuration.

Results of simulations for some parental configurations with two observable alleles are shown in Table 5. Theoretically each of these configurations produces a unique offspring distribution, but some of the offspring distributions have similar class probabilities and the number of offspring has to be increased from 100 to 500 to have a high probability of predicting both parents correctly, or a high power for detecting double-reduction. The last two configurations, (1110, 1220) and (1110, 1120), can be distinguished from other configurations only by the frequency of double-reduction phenotypes, which occur with an overall probability of 0.0131 and 0.005 respectively: hence the lower proportion of correct parental identification.

The application of this approach to dominant markers was also investigated for true parental configurations (1000, 0000), (1100, 0000) and (1000, 1000). Using 100 offspring, the correct configuration was identified for at least 97/100 simulations, except for the third case with no double-reduction, where 89/100 simulations identified the correct parental configuration. However the power to detect double-reduction with $\alpha = 0.1$ was at most 9%, rising to 40% if the number of offspring was increased to 500.

Analysis of experimental data at a simple sequence repeat in potato

This approach is illustrated for a simple sequence repeat (SSR) locus STM003 scored on the parents and 74 offspring of a cross between the advanced SCRI breeding line 12601ab1 and the cultivar Stirling (Bradshaw et al. 1998). Experimental details for scoring the marker phenotype are described by Milbourne et al. (1998).

Table 6 summarises the observed band patterns at the SSR marker locus for the two parents and their offspring classes O_i (i = 1,...8) together with their frequencies. From the parental phenotypes, four genotypes are possible for parent 1 [(1134), (1334), (1344), (1340)] and six genotypes are possible for parent 2 [(2224), (2244), (2444), (2400), (2240), (2440)]. Analysis based upon the offspring phenotypic data shows that the maximum-likelihood estimator of the coefficient of double-reduction is equal to 0.0. In this case there are eight possible configurations of parental genotypes which give rise to all the observed offspring categories. The probabilities for each of these eight configurations are presented in Table 6. The configuration (G_1, G_2) = (1334, 2244) is inferred to be the most-likely for the parental lines with a probability > 0.9999. A goodness-of-fit test ($\chi^2_{df\,=\,7}$ = 6.2, $P > 0.5$) indicates that the observed distribution of offspring band patterns is in good agreement with the expected distribution given the most-likely parental genotypes. If the coefficient of double-reduction is increased from its maximum-likelihood estimate of 0.0, the probability of parental configuration (G_1, G_2) = (1334, 2244) increases further, but the goodness-of-fit statistic worsens.

Discussion

A number of research projects have recently been launched to develop genetic linkage maps in tetraploid crops such as cultivated potato (Meyer et al. 1998, Milbourne et al. 1998), alfalfa (Yu and Pauls 1993), and even in octoploid ones such as sugarcane (Al-Janabi et al. 1993; Da Silva et al. 1995). In the absence of much theory or any software for analysing linkage between codominant markers in a polyploid cross, these maps have been based on single-dose (simplex) dominant markers which segregate in an expected 1:1 ratio in the progeny. To identify and merge homologous chromosomes, double-dose (duplex) markers have been used (e.g. Yu and Pauls 1993; Da Silva et al. 1995) and Hackett et al. (1998) present a simulation study using this approach. A characteristic of this approach is the large standard errors of recombination frequencies between dominant simplex markers linked in repulsion on homologous chromosomes.

The use of codominant molecular markers, such as SSR markers, for linkage mapping in tetraploid species would seem to be a more powerful approach. A prerequisite for estimation of the recombination frequency between two markers is a knowledge of the genotype of each parent at each marker locus. This is straightforward if both parents carry four different alleles which appear as four distinct bands, but in practice this is unusual (Meyer, personal communication). When one or both parents have fewer than four bands, the analysis becomes more complicated. One possible approach is a manual, logical reconstruction of the parental genotypes based on the separate ratios of presence: absence for each allele. If an allele is present in both parents it is necessary to distinguish between ratios of 3 : 1 (simplex*simplex), 11 : 1 (duplex*simplex) or 35 : 1 (duplex*duplex), while if the allele occurs in one parent the expected ratios are 1 : 1 (simplex) or 5 : 1 (duplex). If the locus has no double-reduction, no segregation distortion and few alleles shared between the parents, this approach will frequently be successful. It is essential, however, to compare the observed and expected joint allele frequencies to detect discrepancies due to, for example, multi-locus markers, or a mis-specified parental phenotype due to a faint band. If double-reduction is present, the segregation ratios are modified from those above and a manual reconstruction becomes more difficult. Manual reconstruction of enough loci to construct a linkage map would also be extremely time-consuming.

This paper develops the theory and computational methodology to predict the genotypes of two parental individuals using their phenotypes and the joint segregation information on their progeny's phenotypes observed at a microsatellite marker locus in tetraploid populations. The model allows for the possibility of null alleles. This approach permits the rapid reconstruction of large numbers of loci of any configuration, and gives a maximum-likelihood estimate of the coefficient of double-reduction, and a test of whether this is significantly different from zero. The conditional probabilities of all possible parental genotypes consistent with their phenotype banding patterns are calculated, enabling the confidence in the parental genotype construction to be assessed. A goodness of fit test highlights loci where the offspring data do not fit the expected frequencies, and therefore alternative hypotheses such as multi-locus markers or a mistyped parental banding pattern need to be investigated.

Simulations show that in most cases both the parental genotypes can be correctly recognised with a probability of nearly 1 when the progeny sample size is 100. The presence of a null-allele does not, in general, cause difficulties in predicting the parental genotypes. There are difficulties if two or more parental configurations lead to the same theoretical distribution in the offspring, which may occur when the parents have two observable alleles and produce only two classes of offspring. In this case it is still possible to follow the segregation of one allele and regard it as a dominant marker for linkage analysis.

Double-reduction is a potential problem in analyzing tetrasomic inheritance where quadrivalents are seen at meiosis. The simulation study shows that the mean of the coefficients of double-reduction over 100 simulations was similar to the simulated value. An incorrect inference of double-reduction was rare (at most 3/100 simulations). The power to detect double-reduction varies considerably, depending on the proportion of identical alleles carried and shared by the parents. When the same alleles exist in multiple doses or are shared by the parents, double-reduction involving these alleles will hardly be reflected as a discernible phenotype in the progeny population since zygotes bearing sister-chromatid gametes are less-likely to be distinguished from non-sister-chromatid gametes. As the number of null-alleles in the parental genotypes increases, it plays the same role as multiple identical alleles and thus decreases the power of the double-reduction test.

The simulation study reveals that a satisfactory prediction of tetrasomic segregation can be achieved in most circumstances by using a full-sib progeny size of about 100. However, the simulation study of Hackett et al. (1998) suggested that a population size of 250 is desirable for constructing genetic linkage maps in tetraploid species, and that homologous chromosomes may be difficult to identify with 100 or fewer progeny. A sample size of 250 is realistic for most genome-mapping experiments in plants. The simulation study here did not consider two practical questions: the effects of distorted segregations on the inference of parental genotypes, and the detection of errors in data entry (which might give an offspring configuration which is impossible given the true parental genotype). The effects of these have yet to be investigated, but it seems likely that they will be more severe for small sample sizes.

The marker-phenotype data of two parental genotypes and their offspring at a polymorphic SSR locus were used as a practical example to demonstrate the present theory and method. The first parental phenotype could have arisen from four different genotypes and the second parental phenotype from six genotypes, giving 24 possible parental configurations. Even with a progeny size of 74, the most-likely parental genotypes were predicted with a probability value of > 0.9999, compared to the second largest probability (0.00000316). No evidence of double-reduction was found at this marker locus. The observed and expected frequencies for the classes of offspring agreed well.

This approach is now being applied as the first step of linkage-map construction in tetraploid species. It permits the identification of the parental genotypes, which may then be sorted according to the degree of information carried for linkage analysis. The test of goodness-of-fit provides a useful diagnostic for loci where no model is a satisfactory fit, which can be checked further. When two or more loci are considered, a major difficulty may be the identification of the linkage phase of alleles at different marker loci. This is a far more complicated task than generally imagined, as already demonstrated in Maliepaard et al. (1997) for the diploid case. The computer routines are being expanded to incorporate the joint segregation of pairs of linked loci in different phases, and will form the basis for software to estimate linkage maps in tetraploid species.

Acknowledgements This research was financially supported by a research grant from the UK Biotechnology and Biological Sciences Research Council. Z.W.L. was also supported in part by National Natural Science Foundation of China (# 39993420), and the other authors by the Scottish Office Agriculture, Environment and Fisheries Department. We are grateful for useful discussions with Dr. R. C. Meyer, and for the referees' comments which have improved the format of the paper.

References

Al-Janabi SM, Honeycutt RJ, McClelland M, Sobral BWS (1993) A genetic linkage map of *Saccharum spontaneum* L. 'SES 208'. Genetics 134: 1249–1260

Bailey NTJ (1961) Introduction to the mathematical theory of genetic linkage. Clarendon Press, Oxford

Bradshaw JE, Hackett CA, Meyer RC, Milbourne D, McNicol JW, Phillips MS, Waugh R (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

Briggs FN, Knowles PF (1967) Introduction to plant breeding. Reinhold Publishing Corporation, USA

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC, Sutherland GR (1993) Incidence and origin of "null" alleles in the $(AC)_n$ microsatellite markers. Am J Hum Genet 52: 922—927

Da Silva JAG, Sorrells ME, Burnquist WL, Tanksley SD (1993) RFLP linkage map and genome analysis of *Saccharum spontaneum*. Genome 36: 782–791

Da Silva JAG, Honeycutt RJ, Burnquist WL, Al-Janabi SM, Sorrells ME, Tanskley SD, Sobral BWS (1995) *Saccharum spontaneum* L. 'SES 208' genetic linkage map combining RFLP– and PCR-based markers. Mol Breed 1: 165–179

De Winton D, Haldane JBS (1931) Linkage in the tetraploid *Primula sinensis*. J Genet 24: 121–144

Fisher RA (1947) The theory of linkage in polysomic inheritance. Phil Trans Roy Soc Lond B 233: 55–87

Fisher RA, Mather K (1943) The inheritance of style length in *Lythrum salicaria*. Ann Eugenics 12: 1–23

Fortini P, Barakat R (1980) Genetic algebras for tetraploidy with several loci. J Math Biol 9: 297–304

Grant V (1971) Plant speciation. Columbia University Press, New York, London

Hackett CA, Bradshaw JE, Meyer RC, McNicol JW, Milbourne D, Waugh R (1998) Linkage analysis in tetraploid species: a simulation study. Genet Res 71: 143–154

Lewis WH (1980) Polyploidy: biological relevance. Plenum Press, New York

Maliepaard C, Jansen J, Van Ooijen JW (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 RC, Milbourne D, Hackett CA, Bradshaw JE, McNicol JW, Waugh R (1998) Linkage analysis in tetraploid potato and association of markers with quantitative resistance to late blight (*Phytophthora infestans*). Mol Gen Genet 259: 150–160

Milbourne D, Meyer RC, Collins AJ, Ramsay LD, Gebhardt C, Waugh R (1998) Isolation, characterisation and mapping of simple sequence repeat loci in potato. Mol Gen Genet 259: 233–245

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

Yu KF, Pauls KP (1993) Segregation of random amplified polymorphic DNA markers and strategies for molecular mapping in tetraploid alfalfa. Genome 36: 844–851