Genet. Res., Camb. (1998), **71**, *pp.* 143–154. *With 7 figures.* Printed in the United Kingdom © 1998 Cambridge University Press

Linkage analysis in tetraploid species: a simulation study

C. A. HACKETT[1]*, J. E. BRADSHAW[2], R. C. MEYER[2], J. W. McNICOL[1], D. MILBOURNE[2] AND R. WAUGH[2]

[1] *Biomathematics and Statistics Scotland, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK*
[2] *Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK*

(*Received 26 September 1997 and in revised form 5 December 1997*)

Summary

A simulation study was performed to investigate methods for mapping single-dose (simplex) and double-dose (duplex) markers, and for identification of homologous chromosomes in an autotetraploid species, and to see how the map accuracy depends on the population size. An initial population of 1000 individuals was simulated, with 30 simplex and 10 duplex markers, and recombination fractions and lod scores were calculated between all pairs of markers. These were used to test the feasibility of mapping the simplex and duplex markers simultaneously. Smaller populations, from 500 to 75 individuals, were then simulated, and the estimates of the pairwise recombination fractions and the derived maps were compared with the true map. It was found that the accuracy of the estimates depended strongly on the type of markers involved, with simplex–simplex coupling pairs being most reliable and simplex–simplex repulsion pairs and duplex–duplex pairs in any configuration but coupling being least reliable. Maps can be assembled using recombination fractions and lod scores from pairs of simplex–simplex markers (coupling and repulsion), duplex–simplex (coupling and repulsion) and duplex–duplex (coupling). The agreement between the map order and the true order was good, although the map distance was generally underestimated at small sample sizes.

1. Introduction

Linkage analysis in polyploid species has not advanced at the same rate as in diploid species, due to the difficulties in working with polyploid genotypes. However, early work on linkage in polyploids dates back to the 1930s (De Winton & Haldane, 1931) and Fisher (1947) provided a detailed theoretical description of the estimation of linkage, using combinatorial theory to enumerate all possible genotypes at two or three linked loci, and to express their probabilities in terms of recombination fractions and the probability of double reduction.

The development of DNA markers (RFLPs, RAPDs, AFLPs, SSRs, etc.) led to further developments in the theory of linkage analysis. Ritter *et al.* (1990) described the construction of a linkage map in a cross between heterozygous, diploid parents using single restriction fragments, i.e. a cross of the type AO × OO or AO × AO, where A indicates the presence and O the absence of a particular allele at a locus. Wu *et al.* (1992) extended this method to estimate linkage between single-dose restriction fragments (SDRFs) in auto- and allopolyploids, assuming that chromosomes in an autopolyploid pair at random and that no double reduction gametes are formed. Wu *et al.* (1992) defined an SDRF as being present in one parent P1, absent in the other parent P2, and segregating in a 1:1 ratio in the progeny. In an autotetraploid this defines a simplex × nulliplex cross, AOOO × OOOO. Two such SDRFs may be linked in coupling on the same chromosome, linked in repulsion on two homologous chromosomes, or completely unlinked. However Wu *et al.* (1992) showed that, for autopolyploids, estimates of recombination fractions for markers in coupling have much smaller standard errors than for markers in repulsion. They concluded that large numbers of

* Corresponding author. Tel: +44 (0)1382 562731. Fax: +44 (0)1382 562426. e-mail: chacke@scri.sari.ac.uk.

progeny are required if reliable information on markers in repulsion is to be used to identify groups of homologous chromosomes.

Homologous groups of chromosomes may also be identified using multidose markers. In an autotetraploid species, double-dose markers may segregate in a duplex × nulliplex cross, AAOO × OOOO, giving an expected 5:1 presence:absence ratio in the progeny. Triple-dose (triplex) markers do not segregate in tetraploids, but in species with higher ploidy levels triplex markers may also identify homologies. Da Silva (1993) developed methodology for mapping duplex and triplex markers in an autopolyploid and applied these (Da Silva, 1993; Da Silva *et al*., 1995) to the autooctoploid sugarcane (*Saccharum spontaneum*). Yu & Pauls (1993) examined linkages between a small set of markers scored on a population of tetraploid alfalfa.

The present study was motivated by a new breeding programme at the Scottish Crop Research Institute in the cultivated potato (*Solanum tuberosum* subsp. *tuberosum*), a tetraploid that displays tetrasomic inheritance. The main aim of this programme is to combine quantitative resistances to late blight (*Phytophthora infestans* (Mont.) de Bary) and the white potato cyst nematode (*Globodera pallida* (Stone)) with commercially acceptable tuber yields and quality, if possible using marker-assisted selection. As these resistances have already been introgressed into the cultivated potato, it is advantageous to map at the tetraploid level, thus avoiding time and effort on haploidization and also the problems of making inferences from the diploid to the tetraploid level. The experimental data are presented by Meyer *et al*. (1998).

2. Theory

(i) *Detection and estimation of recombination fractions between markers*

Methodology for the estimation of linkage between simplex markers in autopolyploids has been presented by Wu *et al*. (1992). Da Silva (1993) gave some theory for autooctoploids, but did not include the calculation of the asymptotic variances of the recombination fraction estimators. Yu & Pauls (1993) gave formulae for some estimators for an autotetraploid, but did not explore the relationships between the estimators in much detail. Here we present details of maximum likelihood estimation of the estimators and calculation of the asymptotic variances for an autotetraploid species.

(*a*) *Simplex–simplex linkages*

We consider a population of n progeny from a cross between two parents, P1 and P2, of an autotetraploid species. Chromosomes are assumed to pair at random at meiosis, forming bivalents. Let A and B be two dominant simplex markers (or alternatively alleles of two co-dominant markers), both present in parent P1 and absent in P2, and both segregating in a 1:1 ratio in the progeny. The progeny may be classified into four phenotypic classes – AB, AO, OB and OO – with observed numbers a, b, c and d. If A and B are unlinked, these classes will have equal frequency, and this may be tested by a χ^2 test for independent segregation. A significant result suggests that A and B are linked, either in coupling or in repulsion. If A and B are linked in coupling on one chromosome, that chromosome will pair with one carrying neither marker, and every individual will be informative as to whether a crossover has occurred between A and B. The expected phenotype frequencies are the same as those expected from a backcross in a diploid species, and the recombination fraction is simply the proportion of non-parental progeny. If A and B are linked in repulsion on homologous chromosomes, they have a 1 in 3 chance of being paired together. Therefore, only one-third of the progeny are expected to carry information about crossovers between A and B, considering AB and OO as the recombinant classes in this case. Two-thirds of the chromosome combinations will generate progeny genotypes with an expected 1:1:1:1 distribution of the four genotypic classes, thus also generating 'false recombinants'. Table 1 gives the expected phenotype frequencies for coupling and repulsion linkages. All the phenotypic frequencies in Table 1 were also derived by Yu & Pauls (1993).

The log-likelihood, L, for such phenotypic data is given by

$$L(r) = \text{constant} + a\log(p_{\mathrm{AB}}) + b\log(p_{\mathrm{AO}}) + c\log(p_{\mathrm{OB}}) + d\log(p_{\mathrm{OO}}), \quad (1)$$

where p_{AB}, p_{AO}, p_{OB} and p_{OO} are the expected frequencies of phenotypic classes AB, AO, OB and OO respectively. The recombination fraction, r, is calculated by maximizing the log-likelihood with respect to r. For simplex–simplex linkages we obtain simple formulae:

$$\left.\begin{aligned} &\text{coupling:}\quad \hat{r}_{\mathrm{C}} = (b+c)/n, \\ &\text{repulsion:}\quad \hat{r}_{\mathrm{R}} = [3(a+d)/n] - 1. \end{aligned}\right\} \quad (2)$$

The estimate of the repulsion recombination fraction can be negative due to random variation of the observed phenotype numbers. In this case the maximum of the log-likelihood over the range $[0, 0{\cdot}5]$ is at $r_{\mathrm{R}} = 0$, and this is used as the estimate of r_{R}. This point is discussed later.

The information $I(r)$ is

$$I(r) = -E\left[\frac{d^2L}{dr^2}\right], \quad (3)$$

https://doi.org/10.1017/S0016672398003188 Published online by Cambridge University Press

Table 1. *Expected phenotype frequencies for pairs of simplex (S) and duplex (D) dominant markers linked in coupling and repulsion in an autotetraploid and asymptotic variances of the estimates of the recombination fraction, r*

Linkage type			Phenotypic marker classes in progeny				
Marker A	Marker B	Phase	AB	AO	OB	OO	Variance
S	S	Coupling	$(1-r)/2$	$r/2$	$r/2$	$(1-r)/2$	$\frac{r(1-r)}{n}$
S	S	Repulsion	$(1+r)/6$	$(2-r)/6$	$(2-r)/6$	$(1+r)/6$	$\frac{(1+r)(2-r)}{n}$
D	S	Coupling	$(3-r)/6$	$(2+r)/6$	$r/6$	$(1-r)/6$	$\frac{r(1-r)(2+r)(3-r)}{n(1+r-r^2)}$
D	S	Repulsion	$(2+r)/6$	$(3-r)/6$	$(1-r)/6$	$r/6$	$\frac{r(1-r)(2+r)(3-r)}{n(1+r-r^2)}$
D	D	Coupling	$(5-2r+r^2)/6$	$r(2-r)/6$	$r(2-r)/6$	$(1-2r+r^2)/6$	$\frac{3r(2-r)(5-2r+r^2)}{4n(5-6r+3r^2)}$
D	D	Mixed	$(8+r-r^2)/12$	$(2-r+r^2)/12$	$(2-r+r^2)/12$	$r(1-r)/12$	$\frac{3r(1-r)(2-r+r^2)(8+r-r^2)}{n(4-13r+r^2+24r^3-12r^4)}$
D	D	Repulsion	$(4+r^2)/6$	$(1-r^2)/6$	$(1-r^2)/6$	$r^2/6$	$\frac{3(1+r)(1-r)(4+r^2)}{4n(2+3r^2)}$

where E denotes expected value. The asymptotic variance of the estimator is $1/I(\hat{r})$. These variances are given in the last column of Table 1. Lod scores are calculated as

$$\text{LOD} = L(r = \hat{r}) - L(r = 0{\cdot}5). \tag{4}$$

(*b*) *Duplex–simplex linkages*

Now assume marker A is present in a double dose in parent P1 but absent in P2. A is then expected to be present in five-sixths of the offspring. Linkage between A and a simplex marker B may be detected using a χ^2 test for independent segregation: as before, a significant result indicates linkage. There are two possible situations: either the simplex marker lies on the same chromosome as one of the duplex alleles (coupling) or else it lies on a third homologous chromosome. In both situations we expect two-thirds of the progeny to carry some information about crossovers between the markers. The expected phenotype frequencies are given in Table 1. The maximum likelihood equations may be simplified to a single cubic equation by introducing an intermediate variable x:

$$\left.\begin{aligned} &nx^3+(a-4b-2c-d)\,x^2-(2a-3b+5c\\ &+6d)\,x+6c=0,\\ &\text{coupling:}\quad 0\leqslant x\leqslant 0{\cdot}5;\quad \hat{r}_{\text{C}}=x,\\ &\text{repulsion:}\quad 0{\cdot}5\leqslant x\leqslant 1;\quad \hat{r}_{\text{R}}=1-x, \end{aligned}\right\} \tag{5}$$

but there are no simple explicit formulae for the recombination fractions. Here it is easiest to maximize the log-likelihood numerically. The variance of the estimator is the same for both coupling and repulsion linkages, and is given in Table 1.

(*c*) *Duplex–duplex linkages*

Now let both A and B be duplex markers, present in P1 and absent in P2. If A and B are linked, there are three possible configurations for the four homologous chromosomes: AB/AB/OO/OO (coupling), AB/AO/OB/OO (mixed) or AO/AO/OB/OB (repulsion). In these three situations two-thirds, one-third and two-thirds of the progeny respectively provide information about crossovers. Table 1 gives the expected phenotype frequencies. The maximum likelihood equations simplify to a single quadratic equation for an intermediate variable y:

$$ny^2-(a-4b-4c-3d)\,y-4d=0 \tag{6}$$

with positive solution

$$\left.\begin{aligned} &y=\{(a-4b-4c-3d)+\sqrt{[(a-4b-4c-3d)^2}\\ &\quad +16nd]\}/2n,\\ &\text{coupling:}\quad 0{\cdot}25\leqslant y\leqslant 1;\quad \hat{r}_{\text{C}}=1-\sqrt{y},\\ &\text{repulsion:}\quad 0\leqslant y\leqslant 0{\cdot}25;\quad \hat{r}_{\text{R}}=\sqrt{y},\\ &\text{mixed:}\quad 0\leqslant y\leqslant 0{\cdot}125;\quad \hat{r}_{\text{M}}=[1-\\ &\qquad\qquad\qquad\qquad \sqrt{(1-8y)}]/2. \end{aligned}\right\} \tag{7}$$

https://doi.org/10.1017/S0016672398003188 Published online by Cambridge University Press

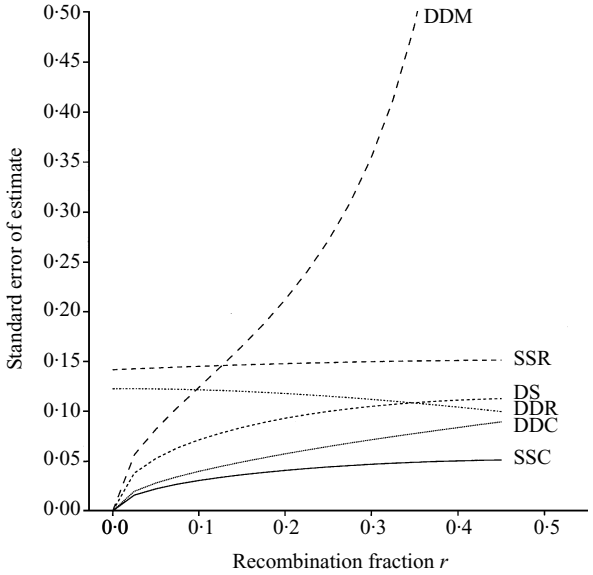

Fig. 1. Standard errors of the estimates of r using a population of size 100. SSC, simplex–simplex coupling; SSR, simplex–simplex repulsion; DS, duplex–simplex (coupling or repulsion); DDC, duplex–duplex coupling; DDM, duplex–duplex mixed; DDR, duplex–duplex repulsion.

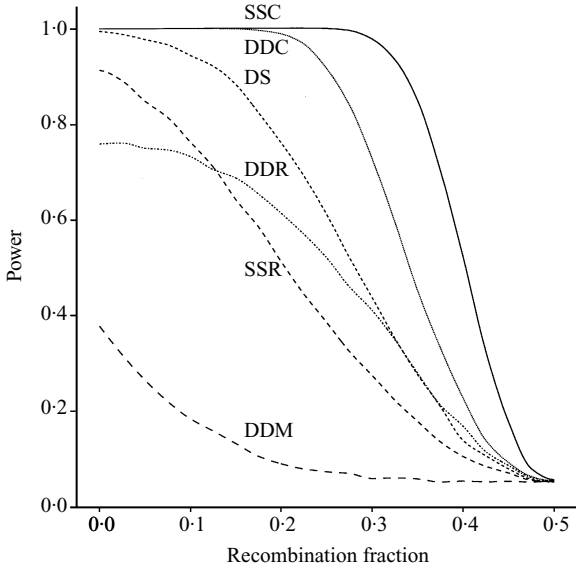

Fig. 2. Power of the test for independent segregation of two markers, for a population of size 100. The y-axis shows the probability of rejection of the hypothesis of unlinked markers (segregating independently) in favour of linkage, given the true recombination fraction on the x-axis. SSC, simplex–simplex coupling; SSR, simplex–simplex repulsion; DS, duplex–simplex (coupling or repulsion); DDC, duplex–duplex coupling; DDM, duplex–duplex mixed; DDR, duplex–duplex repulsion.

Therefore if the solution of the maximum likelihood equation, y, is in the range $[0, 0{\cdot}125]$ there are two possible situations, i.e. the markers may be linked in repulsion or in a mixed configuration, and these cannot be distinguished without reference to further markers. Yu & Pauls (1993) seem to have overlooked this: they imply that the two situations may be separated by means of a large enough population. The variances of the estimators are given in Table 1.

All the expected standard errors, calculated as the square root of the variance, are illustrated in Fig. 1 for a population of 100 progeny. The standard errors associated with duplex markers in a mixed configuration increase very rapidly with the size of the recombination fraction r. The smallest expected standard errors over the range $[0, 0{\cdot}3]$ of r are for simplex–simplex coupling, duplex–duplex coupling and duplex–simplex linkages. Simplex–simplex and duplex–duplex repulsion linkages have much larger expected standard errors. The expected standard errors associated with simplex–simplex repulsion linkages vary very little as r changes, while those for duplex–duplex repulsion linkages decrease slightly as r increases.

We are also interested in the power of the test $r = 0{\cdot}5$ against $r < 0{\cdot}5$ for the different types of linkage, i.e. the probability that the hypothesis $r = 0{\cdot}5$ is rejected when the markers are linked. If $r < 0{\cdot}5$, the χ^2 test for independent segregation of two markers has a non-central χ^2_1 distribution, and the non-centrality parameter λ depends on the true recombination fraction, the population size and the type of linkage (Agresti, 1990). Equations for the non-centrality parameter are:

$$\text{simplex–simplex coupling: } \lambda = n(1-2r)^2 \quad (8)$$

$$\text{simplex–simplex repulsion: } \lambda = n(1-2r)^2/9, \quad (9)$$

$$\text{duplex–simplex: } \lambda = n(1-2r)^2/5, \quad (10)$$

$$\text{duplex–duplex coupling: } \lambda = 11n(3-8r+4r^2)^2/153, \quad (11)$$

$$\text{duplex–duplex mixed: } 19n(1-4r+4r^2)^2/693, \quad (12)$$

$$\text{duplex–duplex repulsion: } \lambda = 11n(1-4r^2)^2/153. \quad (13)$$

The probability that $r = 0{\cdot}5$ is rejected by a test with significance level $\alpha = 0{\cdot}05$ is plotted in Fig. 2 for a population of 100 progeny. These probabilities were obtained from simulated samples of 10000 observations from a non-central χ^2_1 distribution with the desired non-centrality parameter, and so are not perfectly smooth. (Tabulated values of the non-central χ^2_1 distribution are not sufficiently precise to be of use here.) From Fig. 2 we see that the probability of detecting a true linkage of 0·1 varies from more than 0·999 for simplex–simplex and duplex–duplex coupling linkages, to 0·76 for simplex–simplex repulsion linkages and 0·19 for duplex–duplex in a mixed configuration. The probability of detecting a simplex–simplex repulsion linkage decreases particularly rapidly as the true recombination fraction increases.

https://doi.org/10.1017/S0016672398003188 Published online by Cambridge University Press

3. Simulation study

(i) *Generation of tetraploid populations*

(*a*) *Simulation of initial population, POP1000*

An initial simulated population, POP1000, was formed by crossing two parents, P1 and P2. A total of 40 molecular markers were assumed to be segregating, with bands present in P1 and absent in P2. The 40 markers formed six linkage groups: two groups of 18 markers and four isolated markers, unlinked to any others. Each group of 18 markers contained 14 simplex and four duplex markers. Two of the isolated markers were simplex and two were duplex. The first group of 18 markers (G1) was distributed among the four homologous chromosomes of the linkage group as shown in Fig. 3. The second group of 18 markers (G2) had the same pattern across the homologous chromosomes; but the simulated recombination fractions between adjacent markers in the combined map were twice those of the first group.

Gametes of P1 were formed by pairing the four chromosomes of each set into two pairs at random, and then passing one of each pair, possibly after crossovers, to the gamete. The initial population consisted of 1000 progeny, and each was scored for each marker as present or absent. The individual segregation ratios of each marker were tested, using a χ^2 goodness of fit test to compare observed and expected counts. One marker of the 40 was found to deviate from its expected 1:1 ratio, with an observed count of 463 absent, 537 present, $P = 0{\cdot}019$. All other markers followed the expected 1:1 or 5:1 ratio.

(*b*) *Simulation of smaller populations*

We are not often in the fortunate position of being able to conduct mapping studies on a population of 1000 progeny! The population of 1000 lines provided a first test of the feasibility of assembling a map with both simplex and duplex markers linking the homologous chromosomes. Smaller populations were used to investigate the accuracy of this strategy, and where and when errors in mapping arose. Populations of sizes 500, 250, 150, 100 and 75 were simulated by random sampling, with replacement, of individuals from the initial population (POP1000) of 1000. One hundred populations of each size from 500 to 75 were simulated for analysis.

(ii) *Calculation of recombination fractions*

The individual segregation ratios of each marker were checked to classify it as simplex or duplex. Wu *et al.* (1992) showed that a sample size of 75 was adequate

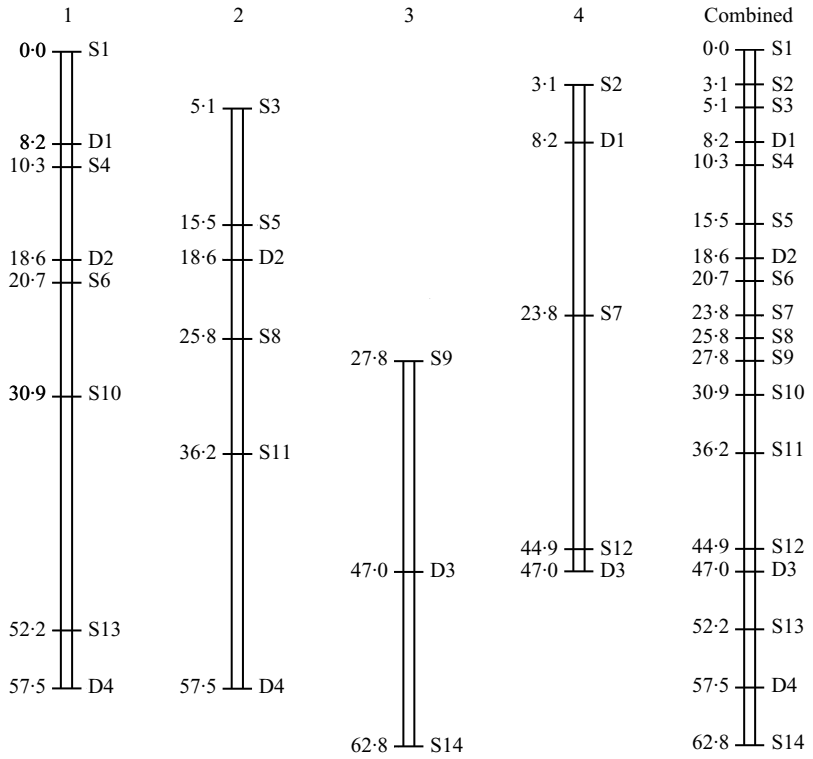

Fig. 3. The simulated arrangement of 18 molecular markers along the four homologous chromosomes of the first linkage group (G1) of parent P1, and on the combined map. S, simplex marker; D, duplex marker. Map positions are given in centimorgans.

https://doi.org/10.1017/S0016672398003188 Published online by Cambridge University Press

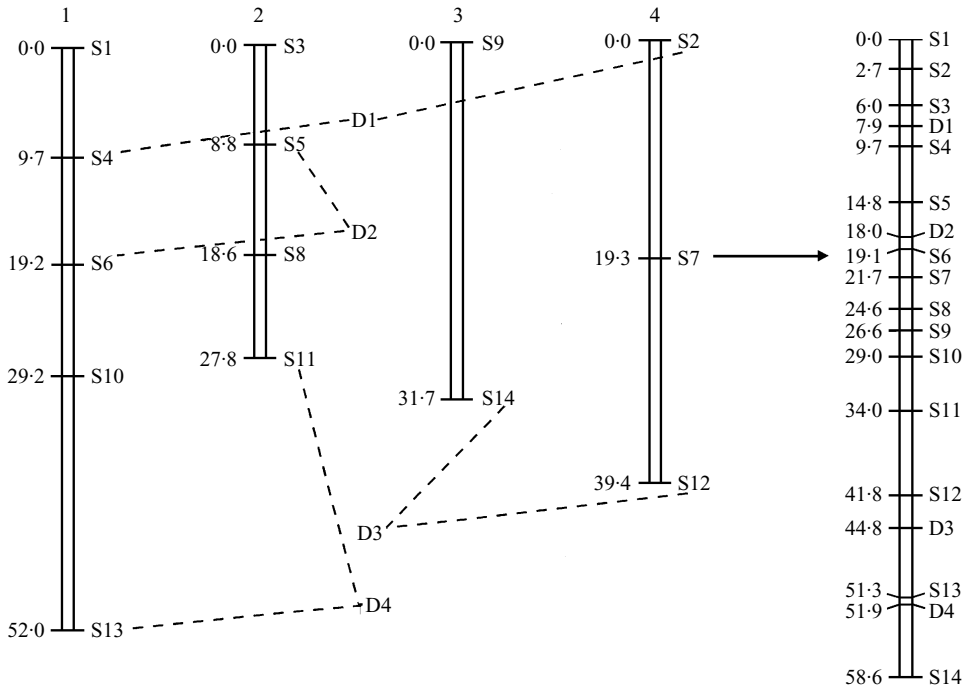

Fig. 4. The estimated maps of the four homologous chromosomes of group G1, obtained from the simplex coupling linkages using 1000 progeny. Dashed lines show the coupling linkages between each duplex marker and the nearest simplex marker on each chromosome. The information from all the duplex–simplex linkages and the duplex–duplex coupling linkages enables these four chromosomes to be merged, giving the complete map on the right.

to distinguish these types with 98% confidence, and each marker here was allocated to the type simulated. Two sets of recombination fractions and lod scores were calculated for POP1000. The first set consisted of recombination fractions between all pairs of simplex markers, r_C, assuming that all the markers were in coupling. Estimates of the recombination fractions greater than 0·5 due to markers in repulsion were truncated to 0·5, and the corresponding lod scores were set to 0·0. This set was used to identify the simplex markers linked on each chromosome.

The second set was used to merge markers from different chromosomes. Repulsion recombination fractions, r_R, were calculated between all pairs of simplex markers, and the overall recombination fraction, r, was taken to be the lower of r_C and r_R. Recombination fractions were calculated for all duplex–simplex pairs by maximizing the log-likelihood numerically. Linkages between all duplex pairs were calculated from (6) and (7). If the data on any duplex–duplex pair were consistent with both a mixed and a repulsion linkage, the recombination fraction assuming a repulsion linkage (r_R) was taken as the estimate. This choice was made because $r_R > r_M$ over the range of recombination fractions [0, 0·333], and because an overestimate of the recombination fraction was thought to be less serious than an underestimate.

All recombination fractions and lod scores were calculated using the statistical package Genstat 5.3 (Genstat 5 Committee, 1993).

(iii) *Mapping the initial population, POP1000, using JoinMap*

The linkage software, JoinMap (Stam & Van Ooijen, 1995), is not able at present to calculate recombination fractions from the raw data for an autotetraploid cross, but it can group and order markers from a set of pairwise recombination fractions, together with the lod scores. The first JoinMap analysis used the recombination fractions and lod scores for all pairs of simplex markers in coupling in POP1000. Ten groups, corresponding to the four chromosomes of group G1, the four chromosomes of G2 and the two unlinked simplex markers, were identified using lod thresholds ranging from 2 to 10. The order of each group corresponded to the order simulated.

The second JoinMap analysis used the complete set of recombination fractions between all pairs of the 40 markers to obtain a combined map, merging the homologous chromosomes. JoinMap identified six linkage groups, the two groups of 18 markers and the four isolated markers, for lod thresholds in the range 4–10. Again the map order corresponded to that simulated, and the total map distance (58·6 cM) compared well with the expected distance of 62·8 cM. Fig. 4 shows the simplex markers and the complete map of group G1, drawn using DrawMap (Van Ooijen, 1994). This map does not enable us to identify which two chromosomes of the four carry each duplex allele, but inspection of the linkage between each

https://doi.org/10.1017/S0016672398003188 Published online by Cambridge University Press

Table 2. *Mean proportion of variance accounted for* (R^2) *by a regression of true recombination fraction on the estimate from different sizes of population, based on 100 simulations of each population size*

	Population size				
Linkage	500	250	150	100	75
SS, coupling	0·96 (0·012)	0·93 (0·023)	0·88 (0·038)	0·82 (0·054)	0·77 (0·074)
SS, repulsion	0·80 (0·045)	0·56 (0·111)	0·30 (0·195)	Zero	Zero
DS, coupling	0·91 (0·020)	0·81 (0·045)	0·71 (0·070)	0·56 (0·103)	0·43 (0·137)
DS, repulsion	0·88 (0·029)	0·77 (0·071)	0·63 (0·107)	0·45 (0·133)	0·26 (0·178)
DD, coupling	0·85 (0·153)	0·62 (0·408)	0·45 (0·462)	Zero	Zero
DD, repulsion	0·40 (0·146)	0·08 (0·349)	Zero	Zero	Zero

The regression line was fixed with slope 1 and intercept 0. The standard deviation of R^2 is given in brackets. 'Zero' indicates that the differences between the true and estimated values were more variable than the true values.

duplex marker and the nearest simplex marker on each chromosome gives this information. Duplex marker D1, for example, is linked in coupling with the markers close to it on chromosomes 1 and 4, and therefore its alleles are located here. The map order of the more distantly linked group G2 also corresponded to that simulated, with a total observed map distance of 125·1 cM compared with an expected distance of 132 cM. The linkage maps were recalculated omitting all duplex–duplex linkages, because of the uncertainty in the linkage phase, and changes in the maps were negligible.

(iv) *Analysis of the smaller populations*

Recombination fractions and lod scores were calculated as described above for each of 100 simulations for each population size from 500 to 75.

(*a*) *Comparison of recombination fraction estimates*

The relationship between the true recombination fractions and the estimated recombination fractions in the smaller populations was investigated by regression analysis, fitting a regression line with slope 1 and intercept 0 to calculate the proportion of variance (R^2) of the true values explained by the estimates. The regression lines were fitted for each type of linkage separately. Estimates of linkages between pairs of markers known to be unlinked were excluded, to avoid a large number of points with recombination fractions close to 0·5. The mean R^2 over the 100 simulations, and the standard deviation, are given in Table 2. The correspondence between the true and estimated values decreased particularly rapidly for simplex and duplex repulsion linkages. The figures for duplex–simplex coupling and repulsion linkages were not significantly different from each other. The results for duplex coupling linkages are included for comparison, but should be treated with caution as each value of R^2 is based on just two estimates.

Fig. 5 shows the estimated recombination fractions plotted against the true values, for one simulation of 250 progeny and one of 75 progeny, and for each type of linkage. These plots emphasize how poor the simplex and duplex repulsion estimates were at low population size. However, the simplex coupling estimates remained quite good. The plot of the simplex repulsion estimates for a population of 75 progeny shows several points with an estimated recombination fraction of zero, while the true recombination fraction takes values up to 0·26. These points correspond to linkages where the estimate was negative, and was replaced by a zero value.

(*b*) *Incidence of negative estimates of simplex repulsion linkages*

For the population of 1000 individuals, one simplex repulsion linkage had a negative estimate among the 435 simplex recombination fractions calculated. Table 3 shows how the mean number of negative estimates increased as the progeny size decreased. The size of the most negative estimate also changed, from a mean of −0·06 for populations of size 500 to a mean of −0·28 for populations of size 75. Table 3 also gives the mean and standard deviation of the mean and largest true recombination fraction for which the estimate was negative for each simulation. For populations of 500 individuals, only pairs of markers with a recombination fraction of up to 0·06 were likely to have a negative estimate, and so a maximum likelihood estimate of 0·0. However, for progenies of size 100, this figure increased to an average of around 0·09, with the largest true recombination fraction having a mean of 0·24. Although we are considering a small proportion of the simplex linkages – on average 20/435 estimates for progenies of 100 or 75 individuals – using an estimate of 0·0 when the true value is above 0·2 may affect the map considerably.

We can understand these results by considering the form of the estimator. Let Y be the number of

https://doi.org/10.1017/S0016672398003188 Published online by Cambridge University Press

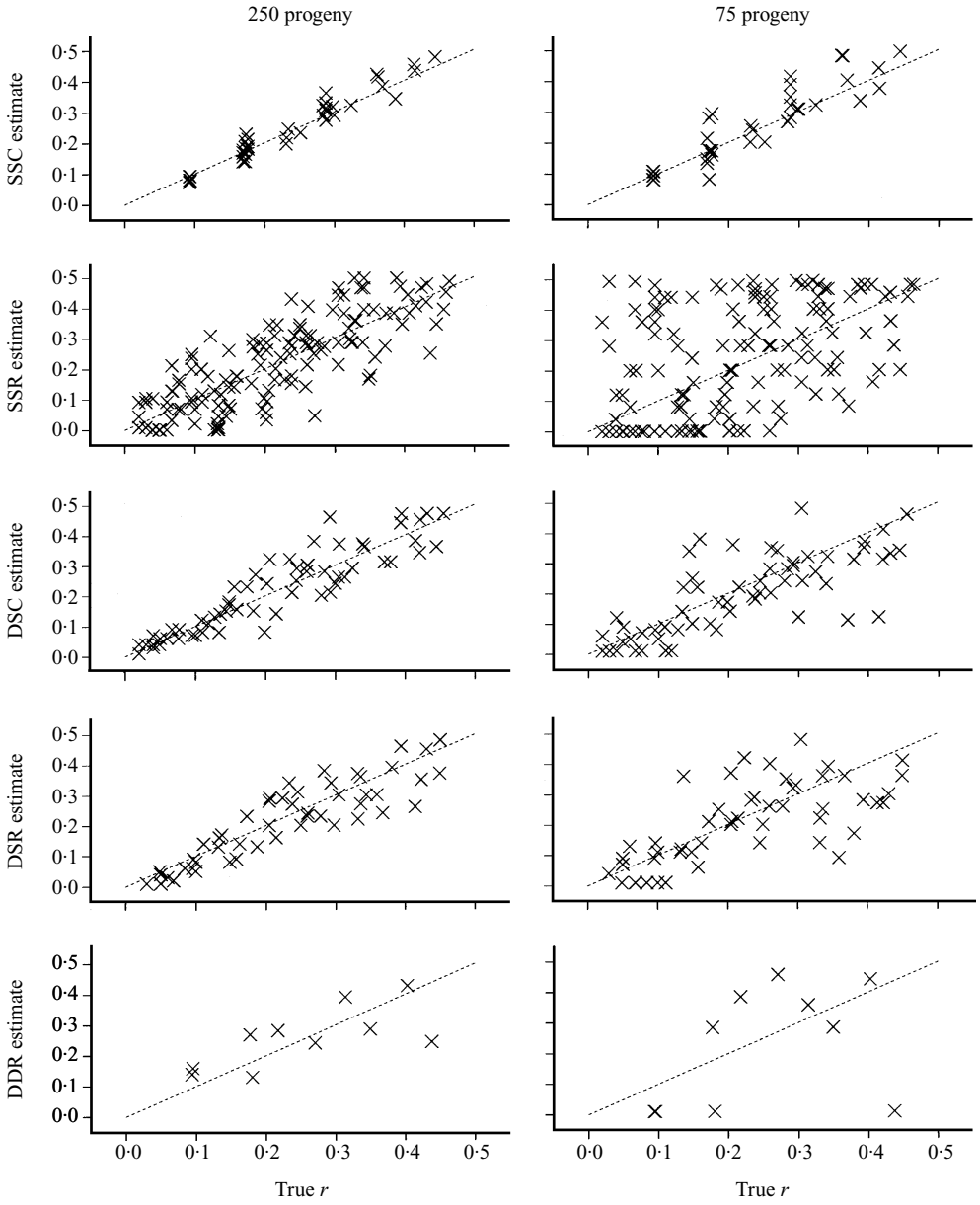

Fig. 5. Comparison of the true and estimated recombination fractions for 250 and 75 progeny, for each type of linkage.

Table 3. *Mean over 100 simulations of the number of simplex repulsion linkages with a negative estimate* (*A*), *the most negative estimate* (*B*), *and the mean* (*C*) *and the largest true recombination fraction* (*D*) *for which the estimate was negative*

	Population size				
	500	250	150	100	75
A	6·0 (2·60)	11·7 (3·46)	13·4 (3·95)	20·7 (5·06)	19·6 (5·28)
B	−0·06 (0·024)	−0·11 (0·037)	−0·16 (0·054)	−0·23 (0·056)	−0·28 (0·085)
C	0·03 (0·013)	0·05 (0·015)	0·07 (0·015)	0·09 (0·017)	0·11 (0·020)
D	0·06 (0·027)	0·13 (0·040)	0·19 (0·051)	0·24 (0·073)	0·35 (0·097)

Standard deviations are given in brackets.

parental types among the n progeny at the two loci in question, i.e. $Y = a + d$ in (2). Then, if the markers are linked in repulsion with a recombination fraction r, Y has a binomial distribution $\mathrm{Bi}(n,(1+r)/3)$. As n is always large here, we can approximate the distribution of Y by a normal distribution with mean $n(1+r)/3$

https://doi.org/10.1017/S0016672398003188 Published online by Cambridge University Press

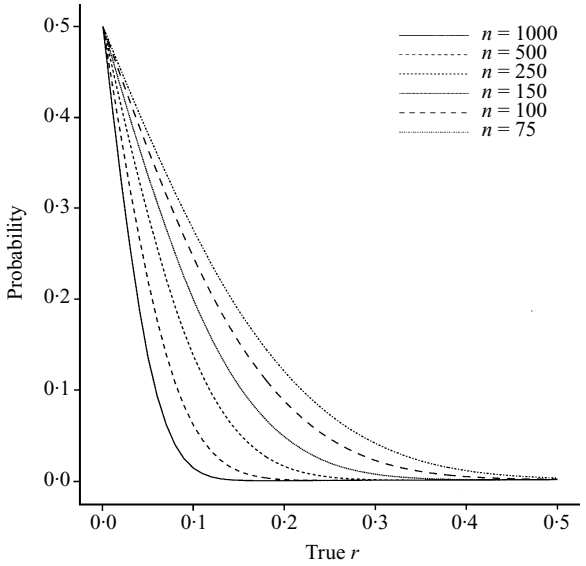

Fig. 6. The theoretical probability of a negative estimate of the simplex repulsion recombination fraction for different population sizes n.

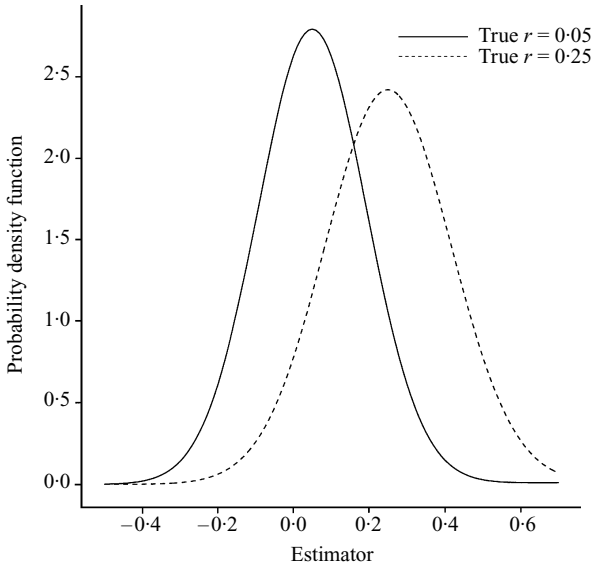

Fig. 7. The theoretical distribution of the simplex repulsion estimator for a population of 100 and true recombination fractions of 0·05 and 0·25.

and variance $n(1+r)(2-r)/9$. The estimator of r_R is $3Y/n-1$, which has a distribution that is approximately normal with mean r and variance $(1+r)(2-r)/n$. The probability that the estimator is negative may be obtained from tables of the normal distribution function, and is shown in Fig. 6 for the different population sizes. We find that as the population size decreases, the probability of a negative estimate of the recombination fraction increases steadily, and so does the range of values of the true recombination fraction for which a negative estimate is possible. For example, recombination fractions of 0·06 have a 10% probability of a negative estimator using 1000 progeny; for a progeny of 100 the corresponding recombination fraction has increased to 0·19.

Fig. 7 compares the theoretical distributions of the estimators for 100 progeny and for true recombination fractions of 0·05 and 0·25. For a true r of 0·05, the estimator is not only more likely to be negative, but likely to have a more negative value than for a true r of 0·25. The actual value of the estimator is ignored under the current method of maximum likelihood estimation if it lies outside the range [0, 0·5]: there is scope here to improve the estimator.

(*c*) *JoinMap analyses*

For each simulated population JoinMap was supplied with the recombination fractions and lod scores between all pairs of simplex markers (using the smaller of the coupling and the repulsion estimates) and those between all simplex–duplex pairs. As the standard errors for duplex–duplex linkages in repulsion or mixed phase were large, and the phase was uncertain, duplex–duplex linkages were excluded. Duplex–duplex coupling linkages are more precise, but this simulation contained no close linkages of this type.

The JoinMap module JMPWG was used to see how the set of pairwise recombination fractions split into groups at different lod thresholds. For populations of 500 progeny the division into the correct six groups (i.e. G1, G2 and four unlinked markers) varied little over a wide range of lod scores. The most common situation was for the correct six groups to be identified for lod thresholds from 3 to 10. At a population size of 250, the correct groups were again identified, although the range of lod thresholds was reduced – between 3 and 6 was the most common range. At a population size of 150, the correct groups were obtained for 84 of 100 simulations, with the most common range of lod thresholds being 2–3. For the smaller populations the groups were more fragmented.

The markers were split up into linkage groups using a lod score of 3·0. Table 4 shows the numbers of simulations with the correct set of 18 markers, with extra markers and with omitted markers for sets G1 and G2. The last line of Table 4 shows the number of simulations where G1 and G2 had merged at a lod score of 3·0. Extra markers in the linkage groups occurred infrequently: they usually resulted in a worsening in the goodness of fit statistic when they were placed on the linkage map and the problem could be resolved in each case by increasing the lod threshold to 4·0, and so removing that marker. Table 4 also shows the mean number of markers placed in each linkage group, averaging over the simulations with no extra markers. For the tightly linked group

https://doi.org/10.1017/S0016672398003188 Published online by Cambridge University Press

Table 4. *The number of simulations with the correct group of 18 markers, with extra markers, and with some markers missing from the linkage group for G1 and G2, using a lod threshold of 3·0. The last line shows the number of simulations where G1 and G2 had merged at a lod score of 3·0. The calculation of the mean and standard deviation (in brackets) of the group size excludes simulations with extra markers*

		Population size				
		500	250	150	100	75
Group G1	> 18	2	0	2	7	7
	18	97	96	95	84	53
	< 18	0	0	0	6	38
	Mean size	18 (0)	18 (0)	18 (0)	17·9 (0·31)	16·7 (2·41)
Group G2	> 18	1	0	0	0	1
	18	98	96	73	11	3
	< 18	0	0	24	86	94
	Mean size	18 (0)	18 (0)	17·6 (0·68)	15·1 (2·13)	11·2 (3·84)
G1 and G2 merged	—	1	4	3	3	2

Table 5. *Comparison of maps from different population sizes*

		Population size				
		500	250	150	100	75
Group G1	Rank correlation	0·996 (0·002)	0·986 (0·017)	0·952 (0·123)	0·907 (0·108)	0·861 (0·189)
	5% point	0·992	0·961	0·893	0·601	0·460
	Shift	2·5 (1·02)	3·7 (1·45)	5·2 (2·68)	6·7 (2·77)	8·1 (3·42)
	Map length	57·3 (2·90)	55·7 (4·61)	51·8 (6·65)	49·8 (7·31)	46·8 (8·14)
	Number of simulations	97	96	95	90	88
Group G2	Rank correlation	0·999 (0·002)	0·995 (0·005)	0·986 (0·012)	0·952 (0·120)	0·937 (0·071)
	5% point	0·994	0·984	0·961	0·886	0·782
	Shift	4·3 (1·17)	5·8 (1·77)	8·3 (2·77)	10·2 (4·56)	11·3 (4·96)
	Map length	124·7 (5·35)	117·3 (7·82)	105·7 (10·65)	92·8 (12·17)	84·8 (11·08)
	Number of simulations	98	96	97	95	61

The rank correlation compares the order of the markers with the true order for that group. Shift gives the mean of the absolute distance of each marker from its map position on the true map (in cM) and map length gives the mean total map length. The true map lengths are 62·8 cM and 132 cM for groups G1 and G2. Means and standard deviations (in brackets) are taken over simulations where the linkage group has 10–18 markers.

G1, the average number of markers placed in the linkage group was 16·7 for a population of 75, and greater for larger populations. For group G2 the average number of markers placed was lower. The markers most frequently omitted from the linkage groups were one or both of the simplex markers on the third chromosome of each set and the duplex marker D4 (see Fig. 3). The simplex markers on the third chromosome are widely separated, with recombination fractions of 0·24 between S9 and S14 in G1 and 0·35 between S24 and S29 in G2, and so their inclusion in the linkage group relies mainly on duplex–simplex linkages. For progenies of 100 and 75 individuals, the markers of sets G1 and G2 tended to fragment due to the separation of one or more groups, each corresponding to a single chromosome.

The linkage groups from each simulation were ordered to form linkage maps, using Haldane's mapping function. Three statistics were used to summarize the linkage map: the Spearman's rank correlation between the true and the estimated marker order, the mean of the absolute distance between the position of each marker on the true map and that on the estimated map, and the total length of the estimated map. Table 5 summarizes the mean and standard deviation of each statistic for simulations where between 10 and 18 markers were placed on the map. Simulations with extra markers were excluded, and so were simulations so fragmented that the largest linkage group had fewer than 10 markers. The lower 5% points of the rank correlations are also included. The mean rank correlation between the true and

https://doi.org/10.1017/S0016672398003188 Published online by Cambridge University Press

estimated orders decreased as the population size decreased, and decreased more rapidly for G1 than for G2. This is reasonable, as the markers of set G1 were more closely linked, and so more likely to exchange positions with a neighbour. The 5%, 1% and 0·1% points of the distribution of the rank correlation for 18 observations are 0·401, 0·550 and 0·692 respectively. For every population size the mean rank correlation was larger than the 0·1% point and even for the worst case, group G1 and a population size of 75, the observed and true orders were significantly correlated at the 5% level for at least 95 of the 100 simulations. We conclude, therefore, that even at the smallest population size considered here the estimated map order contains useful information about the true order. However, as the population size decreased the mean total map length decreased steadily, and the mean difference in position of each marker from its true position increased. This may be due to the effect of simplex repulsion linkages set to zero when the recombination fraction estimate went negative, as zero is certainly an underestimate of the true recombination fraction.

4. Discussion

The aims of the study were to develop a strategy for estimation of a linkage map in an autotetraploid species, and to investigate the effect of population size on the accuracy of the map. The simulations suggest that a population size of at least 150 individuals should be used, and that a larger number – say 250 – would provide a better chance of identifying homologous chromosomes. However, if it is not possible to use as many individuals as this, we found that the simplex coupling data is still reliable, that our mapping strategy is more likely to omit linkages than to find spurious ones, and that the map order is significantly correlated with the true order.

We found the following strategy to be useful in developing a linkage map:

1. *Exploratory analysis* (using Genstat or a similar statistical program)

The first step is to identify the dosage of each marker from the ratio presence : absence. If only one parent carries the marker, the choice is between 1 : 1 and 5 : 1. If both parents carry the marker, then a 3 : 1 ratio suggests a single dose in each parent while a ratio of 11 : 1 or greater suggests at least one parent has a double dose. In this study, the dosage of each marker was known in advance, and was simply confirmed by the ratio presence : absence, but in a real data set we need to test for the dosage and so would prefer to exclude all markers with distorted ratios, at least from the initial map development. We have also found histograms of the number of bands from each parent carried by each progeny to be of interest – this would identify any selfed progeny, and indicated an unusual subset of individuals in the analysis of a real potato data set (Meyer *et al*., 1998).

We then need to calculate recombination fractions and lod scores between pairs of markers. We recommend two such data files: one for simplex coupling linkages only, with recombination fractions greater than 0·5 truncated to 0·5 and the corresponding lod score set to zero, and one including duplex–simplex and simplex repulsion linkages. Duplex–duplex linkages should be included in this second file only if the root of (6) is large enough to indicate a coupling linkage. These files are then ready for analysis using JoinMap.

2. *JoinMap analysis*

A preliminary analysis of the simplex coupling data using JoinMap provides a preliminary order for the markers along each chromosome. The main JoinMap analysis then combines markers from the four homologous chromosomes into a single order. Using a lod threshold of 3·0, we found that markers were seldom placed with the wrong linkage group, and that when this did occur the addition of an incorrect marker to the map caused a sharp jump in the goodness of fit statistic. It is worth changing parameters such as the jump threshold and the triplet threshold to observe the effect on the goodness of fit measure, and also to compare the maps with and without the orders from the simplex coupling analysis as fixed orders. Our simulations showed that the order of markers was usually quite highly correlated with the true marker order, even at the smaller population sizes considered here. However, the length of the map may underestimate the true length of the linkage group, especially if a small population is used.

A problem occurred with simplex repulsion recombination fractions, where it is possible to obtain a negative estimate. In this case the maximum likelihood estimate over the range [0, 0·5] was zero, and this was used as an estimate. This may be an explanation for the shortening of the linkage groups compared with the true map. However, the theory presented in this paper suggests that smaller true recombination fractions may give smaller negative estimates, and therefore that using zero as an estimate, regardless of the size of the negative estimator, is losing some information. We are investigating a Bayesian approach to this problem that would take into account the actual number of parental and recombinant offspring observed.

We thank Scott Chasalow for writing the software to simulate tetraploid individuals. This work was funded by

the Scottish Office Agriculture, Environment and Fisheries Department.

References

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

Da Silva, J. (1993). A methodology for genome mapping of autopolyploids and its application to sugarcane (*Saccharum* spp.) PhD dissertation, Cornell University, Ithaca, NY.

Da Silva, J., Honeycutt, R. J., Burnquist, W., Al-Janabi, S. M., Sorrells, M. E., Tanksley, S. D. & Sobral, B. W. S. (1995). *Saccharum spontaneum* L. 'SES 208' genetic linkage map combining RFLP- and PCR-based markers. *Molecular Breeding* **1**, 165–179.

De Winton, D. & Haldane, J. B. S. (1931). Linkage in the tetraploid *Primula sinensis*. *Journal of Genetics* **24**, 121–144.

Fisher, R. A. (1947). The theory of linkage in polysomic inheritance. *Philosophical Transactions of the Royal Society of London*, *Series B* **233**, 55–87.

Genstat 5 Committee (1993). *Genstat 5 Release 3 Reference Manual*. Oxford: Clarendon Press.

Meyer, R. C., Milbourne, D., Hackett, C. A., Bradshaw, J. E., McNicol, J. W. & Waugh, R. (1998). Linkage analysis in tetraploid potato and associations of markers with quantitative resistance to late blight (*Phytophthora infestans*). *Molecular and General Genetics*, in press.

Ritter, E., Gebhardt, C. & Salamini, F. (1990). Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. *Genetics* **125**, 645–654.

Stam, P. & Van Ooijen, J. W. (1995). *JoinMap version 2.0: Software for the Calculation of Genetic Linkage Maps*. Wageningen: CPRO-DLO.

Van Ooijen, J. W. (1994). DrawMap: a computer program for drawing genetic linkage maps. *Journal of Heredity* **85**, 66.

Wu, K. K., Burnquist, W., Sorrells, M. E., Tew, T. L., Moore, P. H. & Tanksley, S. D. (1992). The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theoretical and Applied Genetics* **83**, 294–300.

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