

# Three-point linkage analysis in crosses of allogamous plant species

M. S. RIDOUT\*, S. TONG†, C. J. VOWDEN‡ AND K. R. TOBUTT

*Horticulture Research International, East Malling, West Malling, Kent ME19 6BJ, UK*

(Received 3 November 1997 and in revised form 27 April 1998)

## Summary

Ritter & Salamini (1996) presented a systematic account of two-point linkage analysis in allogamous diploid plant species. Vowden *et al.* (1995) described an alternative approach that is implemented in a computer program LINKEM. This paper describes how the latter approach has been extended to three-point linkage analysis, and implemented in a new program LINK<sub>3</sub>EM that is available from the authors. The essence of the approach is for the computer program to derive the appropriate form of analysis for a specific cross from its ‘knowledge’ of the most general type of cross that can arise. This avoids the need for programming specific codes for the many different types of cross that can arise. The program allows different locus orderings and parental phases to be compared. The Haldane or Kosambi map functions can be specified, although it is also possible to estimate all three pairwise recombination fractions without any assumed map function.

## 1. Introduction

In a recent paper, Ritter & Salamini (1996) presented a systematic account of two-point linkage analysis in allogamous diploid plant species. They enumerated the possible allelic configurations of the parents, identifying 21 distinct types of cross. For each of these they derived an equation for the maximum likelihood estimator of the recombination fraction (and an explicit formula for the estimator, if one exists) and a formula for the expected information, which can be used to compare the efficiency of different types of cross. Of course many of their formulae have appeared previously in the literature, for example, in Allard (1956), Bailey (1961) and other papers that they cite. But their work provides a general and exhaustive approach that is well suited to implementation within a computer program.

Some time ago, we wrote a Fortran computer program called LINKEM (Vowden *et al.*, 1995) which also does two-point linkage analysis for general crosses of diploid parents. The purpose of the present paper is to describe how the approach used in LINKEM has been

extended to three-point analysis in a more recent program, LINK<sub>3</sub>EM. Three-point linkage analysis is considerably more complicated than two-point analysis, partly because the calculations are more complex but also because issues such as locus ordering and genetic interference become relevant.

The paper is organized as follows. Section 2 outlines briefly the approach to two-point linkage analysis that is used in LINKEM, and contrasts this with the approach of Ritter & Salamini (1996). Section 3 sets out the basic theory of three-point linkage analysis, and indicates how this is implemented in LINK<sub>3</sub>EM, and Section 4 discusses a particular example. The emphasis in this paper is on the methodology used by LINK<sub>3</sub>EM. More detailed information about using the program is provided in the manual (Ridout *et al.*, 1997). The executable version of the program, which runs under MS-DOS on IBM-compatible personal computers, is available on request.

## 2. Two-point linkage analysis

### (i) Two-point linkage analysis in LINKEM

Consider a cross involving two loci, *A* (with alleles  $a_1$ ,  $a_2$ ,  $a_3$  and  $a_4$ ) and *B* (with alleles  $b_1$ ,  $b_2$ ,  $b_3$  and  $b_4$ ). Such a cross has the general form

$$a_1 b_1 / a_2 b_2 \times a_3 b_3 / a_4 b_4. \quad (1)$$

\* Corresponding author.

† Present address: Department of Computer Science, Stanford University, Stanford CA 94305, USA.

‡ Present address: Department of Engineering, University of Warwick, Coventry CV4 7AL, UK.

Table 1. Probabilities of different offspring genotypes resulting from the cross  $a_1 b_1/a_2 b_2 \times a_3 b_3/a_4 b_4$ , when the recombination fraction between loci *A* and *B* is *r*

<i>A</i>	<i>B</i>			
	$b_1 b_3$	$b_1 b_4$	$b_2 b_3$	$b_2 b_4$
$a_1 a_3$	$(1-r)^2/4$	$r(1-r)/4$	$r(1-r)/4$	$r^2/4$
$a_1 a_4$	$r(1-r)/4$	$(1-r)^2/4$	$r^2/4$	$r(1-r)/4$
$a_2 a_3$	$r(1-r)/4$	$r^2/4$	$(1-r)^2/4$	$r(1-r)/4$
$a_2 a_4$	$r^2/4$	$r(1-r)/4$	$r(1-r)/4$	$(1-r)^2/4$

Table 2. Probabilities of different offspring genotypes resulting from the cross  $aw/nx \times aw/nx$ , when the recombination fraction between the two loci is *r*

	<i>ww</i>	<i>wx</i>	<i>xx</i>
$a_-$	$(1-r^2)/4$	$(r^2-r+1)/2$	$r(2-r)/4$
$m$	$r^2/4$	$r(1-r)/2$	$(1-r)^2/4$

Our notation is similar to that of Ritter & Salamini (1996), except that we use lower-case letters and subscripts to denote alleles, writing for example  $a_1$  instead of *A1*, and we do not, at this stage, distinguish null alleles from active alleles. As in their paper, the symbol/indicates the parental phase for linked loci. Thus, for example, in the first (female) parent, alleles  $a_1$  and  $b_1$  lie on one chromosome and alleles  $a_2$  and  $b_2$  on the other.

Let *r* denote the recombination fraction between *A* and *B*. Table 1 shows the probabilities of observing each of the 16 offspring genotype combinations that can result from cross (1). This corresponds to case 21 in Ritter & Salamini (1996). The other cases that they discuss can be obtained by adding together those rows and/or columns of Table 1 for which the offspring genotypes are indistinguishable. For example, consider the cross  $aw/nx \times aw/nx$ , where *n* is a null allele. This is case 10 in their paper. Here, offspring genotypes  $aa$  ( $\equiv a_1 a_3$ ),  $an$  ( $\equiv a_1 a_4$ ) and  $na$  ( $\equiv a_2 a_3$ ) are indistinguishable, as are genotypes  $wx$  ( $\equiv b_1 b_4$ ) and  $xw$  ( $\equiv b_2 b_3$ ). The probabilities of the different offspring genotypes for this cross can therefore be obtained from Table 1 by adding rows 1, 2 and 3 and adding columns 2 and 3. This gives the probabilities shown in Table 2.

LINKEM uses this method to calculate offspring genotype probabilities as functions of the recombination fraction. It stores a copy of Table 1, and determines, from a specification of the cross, which rows and columns, if any, need to be added together. The probabilities in Table 1 are all of the form

$$p = (\beta_2 r^2 + \beta_1 r + \beta_0)/4,$$

where the coefficients  $\beta_0$ ,  $\beta_1$  and  $\beta_2$  are integers. Derived probabilities that result from adding rows and adding columns in Table 1 are, therefore, also of this form, and can be derived with complete accuracy by a computer program using integer arithmetic.

Once the offspring genotype probabilities have been derived, maximum likelihood estimation is straightforward. The log-likelihood function (ignoring terms that do not depend on *r*) is

$$L(r) = \sum n \log(p),$$

where *n* is the number of offspring with a particular genotype, *p* is the corresponding genotype probability, and the summation is over all offspring genotypes. In LINKEM, the log-likelihood function is maximized numerically. After some preliminary checks to see whether the maximum of *L(r)* occurs on the boundary of the parameter space, at  $r = 0$  or at  $r = 0.5$ , LINKEM uses Brent's algorithm (Brent, 1973) to evaluate  $\hat{r}$ , following essentially the implementation in Press *et al.* (1989, section 10.3).

The expected Fisher information, *I(r)*, is then calculated from the formula

$$I(r) = N \sum \frac{1}{p} \left[ \frac{dp}{dr} \right]^2,$$

where *N* is the total number of offspring. This is used to derive the standard error of the estimated recombination fraction, which is the square root of the reciprocal of *I(r)*. LINKEM also derives likelihood-based confidence limits for *r*. Other features of the program are described in the manual (Vowden & Ridout, 1994).

(ii) Comparison with Ritter & Salamini (1996)

Ritter & Salamini (1996) use an efficient code for the calculations relating to the different types of cross that they identify, incorporating an explicit formula for  $\hat{r}$ , if one exists. LINKEM, on the other hand, uses a single algorithm to analyse all types of cross. In particular, it uses numerical optimization to find  $\hat{r}$ , even when an explicit formula is available. This approach is clearly less efficient computationally, but this is of little practical importance, because the program takes only a fraction of a second to complete a single two-point analysis on a personal computer.

(iii) Unknown parental phase

So far, it has been assumed that the parental phase is known. When the phase is not known, LINKEM does a separate analysis for each possible parental phase, of which there may be up to four. Detailed results are presented for the parental phase that gives the largest

value of  $L(\hat{r})$ , but summary results are also given for other plausible parental phases. A parental phase is plausible if it gives rise to a value of  $L(\hat{r})$  that differs from the value under the most likely parental phase by less than a specified constant. This constant can be specified by the user of the program; the default value is 4.0.

### 3. Three-point linkage analysis

This section explains the basic theory of three-point linkage analysis, and describes how this is implemented in the program LINK<sub>3</sub>EM. Many of the ideas we discuss are covered in greater detail by Ott (1991).

#### (i) Derivation of the likelihood function

For three loci,  $A$ ,  $B$  and  $C$ , a general cross has the form

$$a_1 b_1 c_1 / a_2 b_2 c_2 \times a_3 b_3 c_3 / a_4 b_4 c_4.$$

We assume, for now, that the parental phase is known, and that the loci occur in the order  $A-B-C$ . Let  $r_{AB}$  denote the recombination fraction between loci  $A$  and  $B$ , with  $r_{BC}$  and  $r_{AC}$  defined similarly. These recombination fractions are related to the probabilities that the numbers of crossovers between  $A$  and  $B$  and between  $B$  and  $C$  are even (including zero) or odd. We define these probabilities as follows:

Number of crossovers between		
$A$ and $B$	$B$ and $C$	Probability
Even	Even	$\xi_1$
Even	Odd	$\xi_2$
Odd	Even	$\xi_3$
Odd	Odd	$\xi_4$

where  $\xi_1 + \xi_2 + \xi_3 + \xi_4 = 1$ . The recombination fraction  $r_{AB}$  is the probability of an odd number of crossovers between the two loci, irrespective of whether the number of crossovers between loci  $B$  and  $C$  is odd or even. Consequently,  $r_{AB} = \xi_2 + \xi_4$ . Similarly,  $r_{BC} = \xi_2 + \xi_4$ . Recombination between loci  $A$  and  $C$  occurs if there is an odd number of crossovers between  $A$  and  $B$ , or between  $B$  and  $C$ , but not between both, and consequently  $r_{AC} = \xi_1 + \xi_3$ . Conversely, the  $\xi$  values can be calculated from the recombination fractions using the following equations

$$\left. \begin{aligned} \xi_4 &= (r_{AB} + r_{BC} - r_{AC})/2, \\ \xi_2 &= (r_{AB} + r_{AC} - r_{BC})/2, \\ \xi_3 &= (r_{BC} + r_{AC} - r_{AB})/2, \\ \xi_1 &= 1 - \xi_2 - \xi_3 - \xi_4. \end{aligned} \right\} \quad (2)$$

Table 3. Probabilities of different offspring genotypes resulting from the cross  $a_1 b_1 c_1 / a_2 b_2 c_2 \times a_3 b_3 c_3 / a_4 b_4 c_4$  in terms of the probabilities  $\xi_1, \xi_2, \xi_3$  and  $\xi_4$

$A$	$B$	$C$			
		$c_1 c_3$	$c_1 c_4$	$c_2 c_3$	$c_2 c_4$
$a_1 a_3$	$b_1 b_3$	$\xi_1^2/4$	$\xi_1 \xi_2/4$	$\xi_1 \xi_2/4$	$\xi_2^2/4$
	$b_1 b_4$	$\xi_1 \xi_4/4$	$\xi_1 \xi_3/4$	$\xi_2 \xi_4/4$	$\xi_2 \xi_3/4$
	$b_2 b_3$	$\xi_1 \xi_1/4$	$\xi_2 \xi_1/4$	$\xi_1 \xi_3/4$	$\xi_2 \xi_3/4$
	$b_2 b_4$	$\xi_4^2/4$	$\xi_3 \xi_4/4$	$\xi_3 \xi_4/4$	$\xi_3^2/4$
$a_1 a_4$	$b_1 b_3$	$\xi_1 \xi_3/4$	$\xi_1 \xi_4/4$	$\xi_2 \xi_3/4$	$\xi_2 \xi_4/4$
	$b_1 b_4$	$\xi_1 \xi_2/4$	$\xi_1^2/4$	$\xi_2^2/4$	$\xi_1 \xi_2/4$
	$b_2 b_3$	$\xi_3 \xi_4/4$	$\xi_3^2/4$	$\xi_2^2/4$	$\xi_3 \xi_4/4$
	$b_2 b_4$	$\xi_2 \xi_4/4$	$\xi_1 \xi_4/4$	$\xi_2 \xi_3/4$	$\xi_1 \xi_3/4$
$a_2 a_3$	$b_1 b_3$	$\xi_1 \xi_3/4$	$\xi_2 \xi_3/4$	$\xi_1 \xi_4/4$	$\xi_2 \xi_4/4$
	$b_1 b_4$	$\xi_3 \xi_4/4$	$\xi_3^2/4$	$\xi_4^2/4$	$\xi_3 \xi_4/4$
	$b_2 b_3$	$\xi_1 \xi_2/4$	$\xi_1^2/4$	$\xi_1^2/4$	$\xi_1 \xi_2/4$
	$b_2 b_4$	$\xi_2 \xi_1/4$	$\xi_2 \xi_3/4$	$\xi_1 \xi_3/4$	$\xi_1 \xi_4/4$
$a_2 a_4$	$b_1 b_3$	$\xi_2^2/4$	$\xi_3 \xi_1/4$	$\xi_3 \xi_1/4$	$\xi_4^2/4$
	$b_1 b_4$	$\xi_2 \xi_3/4$	$\xi_1 \xi_3/4$	$\xi_2 \xi_4/4$	$\xi_1 \xi_4/4$
	$b_2 b_3$	$\xi_2 \xi_3/4$	$\xi_2 \xi_4/4$	$\xi_1 \xi_3/4$	$\xi_1 \xi_4/4$
	$b_2 b_4$	$\xi_2^2/4$	$\xi_1 \xi_2/4$	$\xi_1 \xi_2/4$	$\xi_1^2/4$

The point of introducing the  $\xi$  values is that they give rise to simple expressions for the probabilities of different offspring genotypes. These are shown in Table 3, which plays the same role in three-point analysis as does Table 1 in two-point linkage analysis.

As in two-point analysis, the offspring genotype probabilities for particular types of cross can be obtained by adding together rows and/or columns of Table 3 for which the offspring genotypes are indistinguishable. The resulting probabilities have the general form

$$\frac{1}{4}(\beta_{11} \xi_1^2 + \beta_{12} \xi_1 \xi_2 + \beta_{13} \xi_1 \xi_3 + \beta_{14} \xi_1 \xi_4 + \beta_{22} \xi_2^2 + \beta_{23} \xi_2 \xi_3 + \beta_{24} \xi_2 \xi_4 + \beta_{33} \xi_3^2 + \beta_{34} \xi_3 \xi_4 + \beta_{44} \xi_4^2), \quad (3)$$

where the  $\beta$  coefficients are integers. The log-likelihood function is again obtained as the sum over all offspring genotypes of the observed frequency multiplied by the natural logarithm of the probability of that genotype.

#### (ii) Calculation of maximum likelihood estimates

In two-point linkage analysis, for example involving loci  $A$  and  $B$ , the recombination fraction  $r_{AB}$  is constrained to lie in the interval  $[0, 0.5]$ . In three-point analysis, this constraint applies to both  $r_{AB}$  and  $r_{BC}$ , whereas the parameter  $r_{AC}$  is constrained to lie in the interval  $[r_{low}, r_{high}]$  where

$$\begin{aligned} r_{low} &= \max(r_{AB}, r_{BC}), \\ r_{high} &= \min(0.5, r_{AB} + r_{BC}). \end{aligned}$$

These constraints assume the locus ordering  $A-B-C$ ; they must be modified in the obvious way for other locus orderings.

Maximum likelihood estimates of parameters are obtained by maximizing the log-likelihood function subject to these constraints. This could be done using a constrained optimization routine. However, in LINK<sub>3</sub>EM the log-likelihood function is first reparametrized in terms of parameters  $\{\phi_1, \phi_2, \phi_3\}$  that can each take *any* real value, and this means that the log-likelihood function can be maximized using an unconstrained optimization routine. The maximum likelihood estimates of the  $\phi$  parameters are transformed to give the maximum likelihood estimates of the  $r$  parameters (maximum likelihood estimates are invariant under parameter transformation). The relationships between the  $\phi$  parameters and the  $r$  parameters are

$$\left. \begin{aligned} r_{AB} &= 0.5/(1 + e^{-\phi_1}), \\ r_{BC} &= 0.5/(1 + e^{-\phi_2}), \\ r_{AC} &= r_{low} + (r_{high} - r_{low})/(1 + e^{-\phi_3}). \end{aligned} \right\} \quad (4)$$

Calculation of the log-likelihood for given values of the  $\phi$  parameters entails transforming to  $r$  parameters, using the equations above, transforming the  $r$  parameters to  $\xi$  parameters, using equations (2), and finally evaluating logarithms of probabilities of the form given by expression (3). The unconstrained optimization uses the Nelder–Mead simplex algorithm (Nelder & Mead, 1965), again following essentially the implementation given in Press *et al.* (1989, section 10.4).

### (iii) Calculation of standard errors and correlations between estimates

In two-point linkage analysis, the variability of the estimated recombination fraction is summarized by its variance, or more commonly by its standard error, which is the square root of the variance. The variance is the reciprocal of the expected information. In three-point linkage analysis, the variability of the estimated recombination fractions is summarized by a *variance matrix*. The diagonal elements of this matrix are the variances of the parameter estimates, and the off-diagonal elements are the covariances of the estimates.

The variance matrix is the inverse of the expected information matrix. This matrix has diagonal elements such as

$$I(r_{AB}, r_{AB}) = N \sum \frac{1}{p} \left[ \frac{\partial p}{\partial r_{AB}} \right]^2$$

and off-diagonal elements such as

$$I(r_{AB}, r_{BC}) = N \sum \frac{1}{p} \left[ \frac{\partial p}{\partial r_{AB}} \right] \left[ \frac{\partial p}{\partial r_{BC}} \right],$$

where, as before,  $N$  is the total number of observations

and the summations are over all possible offspring genotypes.

Derivatives are calculated using formulae such as

$$\frac{\partial p}{\partial r_{AB}} = \sum_{k=1}^4 \left[ \frac{\partial p}{\partial \xi_k} \right] \left[ \frac{\partial \xi_k}{\partial r_{AB}} \right],$$

where  $\partial p / \partial \xi_k$  is obtained by differentiation of equation (3) and  $\partial \xi_k / \partial r_{AB}$  is obtained by differentiation of equations (2). The information matrix is thus derived using *exact* formulae involving the true recombination fractions, rather than by numerical differentiation. This is useful for theoretical efficiency calculations, for example. However, to obtain the variance matrix, the true parameter values are replaced by their estimates and the information matrix is inverted numerically. LINK<sub>3</sub>EM does not display the variance matrix directly. Instead it displays the standard errors of the estimated recombination fractions, and the matrix of correlations between the estimates.

For two-point analysis, LINKEM provides likelihood-based confidence intervals for the estimated recombination fraction, and these are generally more reliable than intervals based on the standard error. In principle, likelihood-based confidence *regions* can be calculated for the estimated recombination fractions in three-point analysis, but these require much more computation and are not available in the current version of LINK<sub>3</sub>EM.

### (iv) Unknown parental phase and/or locus ordering

In two-point linkage analysis there are up to four distinct parental phases, since there are up to two possible chromosomal arrangements in each parent. In three-point analysis there may be as many as 16 distinct parental phases (four in each parent). Additionally, there are three distinct orderings of the loci, giving up to 48 possible phase/ordering combinations. LINK<sub>3</sub>EM allows the user to specify that the parental phase and/or the ordering of the loci is unknown. It then analyses all allowable phase/ordering combinations. Detailed results are given for the phase/ordering that gives the largest log-likelihood, but the program also reports brief summary results for other plausible phase/ordering combinations.

As in two-point analysis, plausible orderings are those that give a maximized value of the log-likelihood function that is not too different from the maximum possible value. This is equivalent to using a likelihood ratio statistic to compare different phase/ordering combinations. However, a comparison of different phase/ordering combinations is, in statistical jargon, a comparison of non-nested models, and there is no simple method of assessing the statistical significance of the likelihood ratio statistic.



(v) *Genetic interference and the coefficient of coincidence*

Recall that  $\xi_4$  has been defined as the probability that there is an odd number of crossovers between *A* and *B* and between *B* and *C* (assuming the locus ordering *A–B–C*). If the occurrence of crossovers between *A* and *B* is independent of their occurrence between *B* and *C* then

$$\xi_4 = r_{AB}r_{BC}.$$

However, departures from independence have been observed in many species. Such departures are quantified by the *coefficient of coincidence*, which is defined as

$$c = \frac{\xi_4}{r_{AB}r_{BC}},$$

where  $c = 1$  when there is independence. It is also possible to express  $\xi_4$ , and hence  $c$ , entirely in terms of the recombination fractions, giving

$$c = \frac{r_{AB} + r_{BC} - r_{AC}}{2r_{AB}r_{BC}}.$$

LINK<sub>3</sub>EM prints the estimated coefficient of coincidence, based on this formula. It also prints an approximate standard error of the estimate, calculated using the general method for calculating the standard error of a function of parameter estimates (Kendall & Stuart, 1976, section 10.6) that is sometimes known as the delta method.

*Genetic interference* is defined as  $1 - c$ . Thus interference is absent if  $c = 1$ , positive if  $c < 1$  and negative if  $c > 1$ . Positive interference, which implies, for example, that the occurrence of an odd number of crossovers between *B* and *C* is less likely if the number of crossovers between *A* and *C* is odd than if it is even, is the more common. If  $c = 0$ , implying that it is *impossible* for there to be an odd number of crossovers in both intervals, interference is said to be *complete*. Because  $c$  is defined as a ratio of probabilities it cannot be negative.

At a more detailed level, genetic interference arises as a result of one or both of two distinct phenomena known as *chiasma* interference and *chromatid* interference (e.g. McPeck, 1996), but the distinction between these is not important for the type of analysis that LINK<sub>3</sub>EM provides.

(vi) *Map functions*

Given three ordered loci, *A–B–C*, the recombination fraction between the two flanking loci is given by the equation

$$r_{AC} = r_{AB} + r_{BC} - 2cr_{AB}r_{BC}, \tag{5}$$

implying that recombination fractions are additive only under complete interference ( $c = 0$ ). Additivity is obtained by transforming recombination fractions to *genetic distances* (or *map distances*), using a *map function*. Different map functions arise from different assumptions about the underlying crossover process. For example, if crossovers occur as a Poisson process, the relationship between map distance ( $x$ ) and recombination fraction ( $r$ ) is given by

$$x = \begin{cases} -\frac{1}{2} \log(1 - 2r) & \text{if } 0 \leq r < 0.5, \\ \infty & \text{if } r = 0.5. \end{cases}$$

This is known as the *Haldane map function* (Haldane, 1919). Another commonly used map function is the *Kosambi map function* (Kosambi, 1944), which is defined as

$$x = \begin{cases} \frac{1}{4} \log \left[ \frac{1 + 2r}{1 - 2r} \right] & \text{if } 0 \leq r < 0.5, \\ \infty & \text{if } r = 0.5. \end{cases}$$

These map functions imply, for ordered loci *A–B–C*, a specific relationship between the three recombination fractions. For the Haldane map function

$$r_{AC} = r_{AB} + r_{BC} - 2r_{AB}r_{BC}, \tag{6}$$

and for the Kosambi map function

$$r_{AC} = \frac{r_{AB} + r_{BC}}{1 + 4r_{AB}r_{BC}}. \tag{7}$$

The Haldane map function is therefore equivalent to assuming  $c = 1$ , whereas for the Kosambi map function  $c$  does not have a fixed value, but instead  $c = 2r_{AC}$ .

LINK<sub>3</sub>EM allows recombination fractions to be estimated assuming either the Haldane or the Kosambi map function. This means that only the recombination fractions  $r_{AB}$  and  $r_{BC}$  are estimated; the value of  $r_{AC}$  is derived using equation (6) or (7). LINK<sub>3</sub>EM also allows the coefficient of coincidence to be fixed, and, given  $r_{AB}$  and  $r_{BC}$ ,  $r_{AC}$  is then calculated using equation (5). This is provided primarily to allow estimation of recombination fractions assuming complete interference, setting  $c = 0$ .

When a map function is specified, and  $r_{AC}$  becomes a known function of  $r_{AB}$  and  $r_{BC}$ , the optimization problem is reduced from three dimensions to two. LINK<sub>3</sub>EM again reparametrizes the problem so as to be able to use the simplex algorithm for unconstrained optimization.

For the fixed  $c$  map function

$$r_{AC} = r_{AB} + r_{BC} - 2cr_{AB}r_{BC}.$$

Given a particular value of  $r_{AB}$  on the interval  $[0, 0.5]$ , one must have

$$r_{BC} \leq \frac{1 - 2r_{AB}}{2(1 - 2cr_{AB})}$$

to ensure that  $r_{AC} \leq 0.5$ . The reparametrization uses parameters  $\{\psi_1, \psi_2\}$  where

$$r_{AB} = \frac{0.5}{(1 + e^{-\psi_1})}$$

$$r_{BC} = \frac{0.5(1 - 2r_{AB})}{(1 - 2cr_{AB})(1 + e^{-\psi_2})}$$

In the special case of the Haldane map function ( $c = 1$ ),  $r_{AB}$  and  $r_{BC}$  may vary freely and independently on the interval  $[0, 0.5]$ . This is also true for the Kosambi map function.

The information matrix becomes  $2 \times 2$ , rather than  $3 \times 3$ , when a map function is assumed. It is calculated in the same way as before, but formulae for the derivatives of the  $\xi$  parameters with respect to the recombination fractions are now more complex. For example, with the Kosambi map function the relationship

$$\xi_4 = (r_{AB} + r_{BC} - r_{AC})/2$$

becomes

$$\xi_4 = \frac{1}{2} \left[ r_{AB} + r_{BC} - \frac{r_{AB} + r_{BC}}{1 + 4r_{AB}r_{BC}} \right]$$

$$= \frac{2r_{AB}r_{BC}(r_{AB} + r_{BC})}{1 + 4r_{AB}r_{BC}}$$

Inversion of the information matrix gives the variance matrix for two of the recombination fractions. The standard error of the remaining recombination fraction is obtained by the delta method.

It is important to emphasize that many of the formulae in the previous two subsections assume the locus ordering  $A-B-C$ . Analogous, but different, formulae apply for other locus orderings.

#### 4. An example of three-point linkage analysis

##### (i) Parental phase and locus ordering assumed known

This section illustrates the output produced by LINK<sub>3</sub>EM, using some artificial data. We use artificial data rather than real data because it enables us to address several issues with a single example. Consider a cross involving three loci, *Loc1*, *Loc2* and *Loc3*, with the following parental genotypes:

$$Loc1: ab \times ab \quad Loc2: uv \times uv \quad Loc3: yz \times yz.$$

We assume that the three loci lie on the same chromosome. For now, we also assume that the locus ordering is *Loc1-Loc2-Loc3* and that the parental phase is known to be

$$auy/bvz \times auy/buz.$$

For this specific example we have used a notation that is simpler than was used in the previous section.

Table 4 gives data for 60 progeny from this cross, showing the numbers of individuals with different

combinations of genotypes at the three loci. For this cross, 18 distinct offspring genotype combinations can occur.

The first column of Table 5 gives selected output from running LINK<sub>3</sub>EM with these data, including details about the locus ordering, the estimated recombination fractions, their standard errors, and the correlations between them. In this example, the estimates are not particularly highly correlated. The output also includes the estimated coefficient of coincidence and its standard error, and a goodness of fit test. The goodness of fit statistic, the deviance, is calculated from the table of observed and fitted values (Table 4) by the formula

$$deviance = 2 \sum observed * \log_e \left( \frac{observed}{fitted} \right),$$

where the sum is over all entries in the table, excluding any cells in which the observed frequency is zero.

When the assumed model is correct, the deviance follows, approximately, a chi-squared distribution. The number of degrees of freedom is 4 less than the number of cells in the table, because the total number of observations is fixed and three parameters have been estimated. The  $\chi^2$  approximation becomes more reliable as the number of offspring increases. The deviance is related directly to the log-likelihood function, and indeed, maximizing the log-likelihood function is equivalent to minimizing the deviance. In this example the deviance does not indicate any significant lack of fit.

The estimates of the three recombination fractions in the first column of Table 5 may be compared with the estimates obtained by doing a separate two-point analysis for each of the three pairs of loci. The two-point estimates and their standard errors are:

Table 4. Hypothetical offspring data from the cross *auy/bvz* × *auy/buz* showing the number of offspring with each of the 18 possible combinations of genotypes. There are 60 offspring in total. Figures in brackets are fitted values based on the estimated recombination fractions given in the first column of Table 5

Loc1	Loc2	Loc3		
		yy	yz	zz
aa	uu	11 (10.8)	2 (2.3)	0 (0.1)
	uw	0 (0.5)	1 (1.2)	0 (0.2)
ab	uu	2 (2.8)	12 (11.5)	1 (0.7)
	uw	0 (0.7)	12 (11.5)	7 (2.8)
bb	uu	0 (0.2)	0 (1.2)	1 (0.5)
	uw	1 (0.1)	0 (2.3)	10 (10.8)

Table 5. Selected output from analysing the data in Table 4 using *LINK<sub>3</sub>EM*, assuming different map functions

	Map function			
	None	Haldane	Kosambi	Complete interference
Locus ordering	<i>Loc1-Loc2-Loc3</i>	<i>Loc1-Loc2-Loc3</i>	<i>Loc1-Loc2-Loc3</i>	<i>Loc3-Loc1-Loc2</i>
Parental phase	<i>auy/bvz</i> × <i>auy/buz</i>	<i>auy/bvz</i> × <i>auy/buz</i>	<i>auy/bvz</i> × <i>auy/buz</i>	<i>yau/zbv</i> × <i>yau/zbu</i>
Estimated recombination fractions <sup>a</sup>				
Between <i>Loc1</i> and <i>Loc2</i> ( $r_1$ )	0.12 (0.051)	0.098 (0.0388)	0.099 (0.0395)	0.054 (0.0409)
Between <i>Loc2</i> and <i>Loc3</i> ( $r_2$ )	0.081 (0.0462)	0.066 (0.0359)	0.066 (0.0358)	0.19 (0.041)
Between <i>Loc1</i> and <i>Loc3</i> ( $r_3$ )	0.13 (0.034)	0.15 (0.035)	0.16 (0.037)	0.13 (0.034)
Correlations between estimates				
$r_1$ and $r_2$	0.26	-0.38	-0.45	-
$r_1$ and $r_3$	0.41	-	-	-0.05
$r_2$ and $r_3$	0.22	-	-	-
Coefficient of coincidence <sup>a</sup>	3.49 (1.62)	1 (fixed)	0.32 (fixed by map function)	0 (fixed)
Goodness of fit statistic	18.50 (14 d.f.; $P = 0.18$ )	20.49 (15 d.f.; $P = 0.15$ )	22.79 (15 d.f.; $P = 0.09$ )	27.47 (15 d.f.; $P = 0.03$ )

<sup>a</sup> SEs in parentheses.

Table 6. Standard error of  $r_{12}$  from three-point analysis as a percentage of the standard error from two-point analysis, for different values of  $r_{12}$  and  $r_{23}$

$r_{12}$	$r_{23}$				
	0.02	0.05	0.10	0.25	0.50
0.02	71.8	81.0	88.3	96.5	100.0
0.05	65.5	73.8	81.6	93.8	100.0
0.10	63.0	70.4	78.2	92.3	100.0
0.25	66.4	73.5	81.7	95.1	100.0
0.50	77.9	85.1	92.2	99.1	100.0

Between *Loc1* and *Loc2*:  $r_{12} = 0.077$  (SE = 0.0487),  
 Between *Loc2* and *Loc3*:  $r_{23} = 0.091$  (SE = 0.0525),  
 Between *Loc1* and *Loc3*:  $r_{13} = 0.13$  (SE = 0.034).

Two- and three-point analyses give the same estimated recombination fraction and standard error for *Loc1* and *Loc3*, to the number of decimal places displayed (0.13, 0.034). For *Loc2* and *Loc3* the estimated recombination fraction is similar in both analyses (0.091 vs 0.081), but the standard error is larger from the two-point analysis (0.0525 vs 0.0462). For *Loc1* and *Loc2* the estimated recombination fraction is rather different in the two analyses (0.077 vs 0.012). Although the difference between the estimates is not large in relation to their standard errors, it might be large enough to affect the outcome of a mapping program.

For *Loc1* and *Loc2*, the standard error of the estimated recombination fraction is slightly smaller in the two-point analysis (0.0487 vs 0.051), but this may be misleading, given the difference in the estimates produced by two- and three-point analysis, because the standard error of an estimated recombination

fraction is usually strongly dependent on the estimate itself. Table 6 presents a fairer comparison of two- and three-point analysis in this example. It shows the standard error from three-point analysis as a percentage of the standard error from two-point analysis, for different true values of the recombination fraction  $r_{12}$ . This ratio depends also on the true values of the other two recombination fractions,  $r_{23}$  and  $r_{13}$ . Different values of  $r_{23}$  are shown in Table 6, and, for convenience, we have assumed that  $r_{13}$  is related to  $r_{12}$  and  $r_{23}$  by the Kosambi formula (cf. equation 7)

$$r_{13} = \frac{r_{12} + r_{23}}{1 + 4r_{12}r_{23}}$$

Table 6 shows that three-point analysis is generally preferable to two-point analysis. The advantage of three-point analysis for estimating  $r_{12}$  decreases as the strength of linkage between *Loc2* and *Loc3* decreases ( $r_{23}$  increases), and two-point and three-point analysis give the same standard error when *Loc2* and *Loc3* are unlinked ( $r_{23} = 0.5$ ).

The increased efficiency of three-point analysis comes about mostly because it is able to glean some information about linkage from the male parent. In two-point analysis the male parent provides no information about  $r_{12}$ , because *Loc2* is homozygous in this parent. But suppose that *Loc2* and *Loc3* are in fact closely linked, so that recombination between these loci occurs only infrequently. Then if the haplotype from this parent is *auy* or *buz*, it is likely that recombination has *not* occurred between *Loc1* and *Loc2*, because recombination would then also have to have occurred between *Loc2* and *Loc3*. Conversely, if the haplotype is *auz* or *buy* it is likely that recombination between *Loc1* and *Loc2* has

occurred. One extreme case is when  $r_{23} = 0$ , so that recombination between *Loc2* and *Loc3* never occurs. Conversely, if  $r_{23} = 0.5$ , so that *Loc2* and *Loc3* segregate independently, three-point analysis provides no additional information, as is reflected by the values of 100% in the final column of Table 6.

(ii) *Unknown parental phase and/or locus ordering*

So far, we have assumed that the parental phase and the locus ordering are known. If the example is re-run with phase and ordering specified as unknown it turns out that the most likely phase/ordering combination is the one used previously, with deviance 18.50. However, with the same parental phase, the ordering *Loc2–Loc3–Loc1* gives deviance 18.56 and the ordering *Loc3–Loc1–Loc2* gives deviance 19.53, and these are flagged as plausible alternatives by LINK<sub>3</sub>EM (with the default setting for plausibility). The next largest value of the deviance, arising from a different parental phase, is 47.16 – much larger than the values obtained previously. The appropriate conclusions are that the three loci are linked, that the parental phase is clearly identified, but that the data give little useful information about the ordering of the three loci on the chromosome.

(iii) *Using map functions*

Referring back to the first column of Table 5 we see that with no assumed map function the estimated coefficient of coincidence is 3.49 (SE = 1.62), indicating negative interference. To test whether this constitutes a significant departure from no interference ( $c = 1$ ) we can calculate the test statistic

$$z = \frac{3.49 - 1}{1.62} = 1.54.$$

If there is no interference,  $z$  has approximately a standard normal distribution. As positive and negative interference can occur, a two-tailed test is appropriate here, and this gives a  $P$  value of 0.12. Thus the evidence for negative interference is not very strong.

We can analyse the data assuming no interference by specifying the Haldane map function, and this gives the output shown in the second column of Table 5. The most likely locus ordering is again *Loc1–Loc2–Loc3*. The estimated recombination fractions  $r_{12}$  and  $r_{23}$  are slightly smaller than before. However, even these smaller estimates imply a slightly larger estimate of  $r_{13}$  than in the unrestricted analysis. The deviance is 20.49 on 15 d.f. whereas in the unrestricted model it was 18.50 on 14 d.f. The increase, 1.99, may be compared with  $\chi^2$  tables on 1 d.f. to obtain an alternative test for interference that is likely to be more accurate than the  $z$ -test given above. In this example it gives a  $P$  value of 0.16, implying even less

evidence of interference. The ordering *Loc3–Loc1–Loc2* is less plausible, as an alternative to *Loc1–Loc2–Loc3*, under the Haldane map function than it was when no map function was assumed.

The third column of Table 5 shows the output if we assume the Kosambi map function instead. Again the most likely ordering is *Loc1–Loc2–Loc3*. The estimated recombination fractions, and their standard errors, are very similar to those obtained with the Haldane map function. However, the deviance is larger, indicating that this map function does not give such a good fit. Unfortunately, there is no simple test for comparing the Haldane and Kosambi map functions directly, since this again involves the comparison of non-nested models. However, we can compare the Kosambi map function with the unrestricted model by comparing the difference in deviance, which is 4.29 (= 22.79 – 18.50), with the  $\chi^2$  distribution on 1 d.f. This gives a  $P$  value of 0.04. Thus the fit of the Kosambi model is significantly worse than the fit of the unrestricted model.

We could reach a similar conclusion by using a  $z$ -test to see whether the value of  $c$  obtained under the Kosambi map function, 0.32, appears reasonable, based on the unrestricted analysis. This gives

$$z = \frac{3.49 - 0.32}{1.62} = 1.96,$$

implying a similar  $P$  value of 0.05.

It is clear from this that assuming complete interference ( $c = 0$ ) will give an even less satisfactory fit. However, for illustration we show the results in the final column of Table 5. This time the results are quite different. The most likely parental phase is as before, but the most likely ordering is now *Loc3–Loc1–Loc2*. In fact, assuming complete interference, this is the only ordering possible with that phase. The ordering *Loc1–Loc2–Loc3*, for example, is impossible because the probability of genotypes  $\{aa, uw, yy\}$  and  $\{bb, uu, zz\}$  is zero with complete interference, but in fact one individual has the latter genotype.

(iv) *Crosses in which each parent is homozygous at one locus*

Suppose that we modify the cross that we have been studying to

$$aay/avz \times auy/bvy.$$

The female parent provides information about  $r_{23}$ , but not about  $r_{12}$  or  $r_{13}$ . Similarly, the male parent provides information about  $r_{12}$ , but not about  $r_{23}$  or  $r_{13}$ . Since neither parent provides any information about  $r_{13}$ , a full three-point analysis is impossible. However, analysis is possible if we are prepared to



assume one of the map functions, since  $r_{13}$  is then a function of  $r_{12}$  and  $r_{23}$ . More generally, given one of the map functions any of the three recombination fractions can be calculated if the other two are known. LINK<sub>3</sub>EM is therefore able to analyse crosses of this type, provided that a map function is specified.

## 5. Discussion

Ritter & Salamini (1996) identified 21 distinct types of cross that need to be considered for an exhaustive treatment of two-point linkage analysis. We have not attempted to enumerate the possibilities for three-point analysis, but the number is clearly much greater than in two-point analysis. As a result, it would be very cumbersome to list all the possible formulae, and it is preferable to use a computer program such as LINK<sub>3</sub>EM to generate formulae as and when they are needed.

Of course, the advantage of not having to derive and program many special formulae is offset, to some extent, by a loss of computational efficiency. We argued earlier that in two-point analysis this loss of efficiency is of little practical importance. This is not entirely true of three-point analysis. A single three-point analysis in LINK<sub>3</sub>EM, if it involves looking at all 48 combinations of parental phase and locus ordering, can take a few seconds of computing time. This could doubtless be reduced by improving the numerical aspects of the program, since this is an area to which we gave low priority in developing this version of LINK<sub>3</sub>EM. For example, the simplex algorithm that we have used for optimization is robust but not very efficient.

Potentially, three-point linkage analysis offers two advantages over two-point analysis. First, and most importantly, three-point analysis allows the likelihood of different orderings of the three loci to be assessed directly. This is important in building genetic maps. Secondly, three-point analysis may provide more precise estimates of recombination fractions. This was certainly the case in the example of Section 3. A referee has suggested another illuminating example. Consider the cross  $AvX/aVx \times AVX/avx$ , with the following offspring genotypes: 28  $\{A-, V-, X-\}$ , 4  $\{A-, V-, xx\}$ , 12  $\{A-, vv, X-\}$ , 3  $\{A-, vv, xx\}$ , 1  $\{aa, V-, X-\}$ , 8  $\{aa, V-, xx\}$ , 2  $\{aa, vv, X-\}$ , 2  $\{aa, vv, xx\}$ . Two-point linkage analysis gives the following estimated recombination fractions, with standard errors in brackets:  $r_{12} = 0.38$  (0.386),  $r_{13} = 0.18$  (0.056) and  $r_{23} = 0.39$  (0.418). Only the first and third loci appear to be linked. However, three-point analysis based on this locus ordering gives the following estimates:  $r_{12} = 0.20$  (0.130),  $r_{13} = 0.20$  (0.059) and  $r_{23} = 0.20$  (0.130). There is little change to  $r_{12}$  or its standard error, but this analysis suggests that

all three loci may be linked. Two-point analysis involving the second locus is not very informative in this example, because in the female parent the dominant allele for the second locus is coupled with the recessive alleles at the other two loci.

Lathrop *et al.* (1985) did similar efficiency calculations and found generally modest improvements in efficiency from using three-point analysis, though much larger improvements resulted in special circumstances that are outside the scope of the present paper, such as when penetrance was incomplete. In another study, Thompson (1984) also found small improvements in efficiency for three-point analysis compared with two-point analysis, but considered that the main benefit from three-point analysis was the improved information about locus ordering.

There is one complication that can arise with three-point analysis that does not arise with two point analysis. Suppose that four loci are believed to lie in the order  $A-B-C-D$ . Three-point analysis of the triples  $A-B-C$  and  $B-C-D$  will usually give rise to two different estimates of the recombination fraction  $r_{BC}$ . The best way of combining these estimates is to take a weighted mean, with the weights being the reciprocals of the variances of the two separate estimates. Often, the resulting estimate will be similar to the two-point estimate, but when the two differ, the combined three-point estimate should be the more reliable.

Currently, LINK<sub>3</sub>EM does not provide any facilities for testing for linkage. Our assumption is that data would first be run through a two-point analysis program, such as LINKEM, to determine linkage groups. The emphasis in the three-point program is, therefore, on estimation rather than testing. However, the example given earlier in the Discussion, based on the cross  $AvX/aVx \times AVX/avx$ , indicates that three-point analysis may detect linkages which are not apparent from two-point analysis. It may therefore be useful to incorporate linkage tests into LINK<sub>3</sub>EM in the future. Ott (1991, section 6.1) discusses various tests that are available.

It is worth emphasizing that the program is quite general. It is not necessary to make any assumptions about parental phase (though this is something that can often be deduced from two-point analysis), locus ordering or interference. There are, of course, numerous programs available for multipoint linkage analysis that allow many more than three loci to be analysed simultaneously. However, most of these assume the Haldane map function, which greatly simplifies the calculations. For example, Stam (1993) points out that estimation of recombination fractions in the program MAPMAKER (Lander *et al.*, 1987) is based on the Haldane map function even when the user chooses the Kosambi map function; the Kosambi function is used only to convert the estimated

recombination fractions to map distances. Stam's own program JOINMAP (Stam & Van Ooijen, 1995), which is available commercially, correctly calculates log-likelihoods for three-point analysis (but not general multipoint analysis) with either the Haldane or the Kosambi functions. However, it is not possible, in JOINMAP, to estimate all three recombination fractions without any assumption about interference.

Experimental studies of various organisms have shown that genetic interference usually exists, but the Haldane map function is often used nonetheless, primarily for its computational convenience. However, an additional justification is sometimes used, particularly in human genetics, where experimental data are not available. It is argued that the level of interference is very difficult to estimate from the available data, resulting in very imprecise estimators. Estimates of recombination that ignore interference can be biased, but may have smaller mean square error than estimates that rely on an estimated level of interference. See Ott (1991, pp. 135–136) for further discussion. It is not clear, however, to what extent this argument is appropriate for experimental organisms, particularly plants, where reasonably large numbers of offspring are often available from a single set of parents.

Extending LINK<sub>3</sub>EM to provide more general multipoint linkage analysis would clearly involve an increase in the complexity of the calculations. But there is another difficulty that can be illustrated by considering four-point analysis. In generalizing equation (2) there would be eight  $\xi$  probabilities, which would sum to one, leaving seven unknowns. On the other hand there would be only six pairwise recombination fractions. Thus it is no longer possible to calculate the  $\xi$  values uniquely from the recombination fractions (e.g. Owen, 1950). It is instead necessary to introduce an additional *generalized* recombination fraction to specify the probability that there are, in total, an odd number of crossovers between the first and second loci and the third and fourth loci (Karlin & Liberman, 1979). Map functions simplify the analysis because all the recombination fractions, including the generalized recombination fraction, can be calculated from the pairwise recombination fractions between adjacent loci. However, for many map functions this can lead to negative values for the  $\xi$  probabilities when there are more than three loci (Liberman & Karlin, 1984). Map functions for which this cannot occur, including the Haldane map function, are said to be multilocus feasible. But most commonly used map functions, including the Kosambi function, are not multilocus feasible. Recently, Speed (1996) has questioned whether one should necessarily consider map functions that are not multilocus feasible to be invalid, but the point we wish to make here is that moving beyond three-point

analysis introduces additional complexities that are not purely computational.

This work was funded in part by European Union grant AIR3-CT93-1585. We are grateful to the referees for their comments on an earlier draft of this paper.

## References

- Allard, R. W. (1956). Formulas and tables to facilitate the calculation of recombination values in heredity. *Hilgardia* **24**, 235–278.
- Bailey, N. T. J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford: Clarendon Press.
- Brent, R. P. (1973). *Algorithms for Minimization Without Derivatives*. New York: Prentice-Hall.
- Haldane, J. B. S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**, 299–309.
- Karlin, S. & Liberman, U. (1979). A natural class of multilocus recombination processes and related measures of crossover interference. *Advances in Applied Probability* **11**, 479–501.
- Kendall, M. G. & Stuart, A. (1976). *The Advanced Theory of Statistics*, vol. I, 4th edn. High Wycombe: Charles Griffin.
- Kosambi, D. D. (1944). The estimation of map distances from recombination values. *Annals of Eugenics* **12**, 172–175.
- Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E. & Newburg, L. (1987). Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**, 174–181.
- Lathrop, G. M., Lalouel, J. M., Julier, C. & Ott, J. (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American Journal of Human Genetics* **37**, 482–498.
- Liberman, U. & Karlin, S. (1984). Theoretical models of genetic map functions. *Theoretical Population Biology* **25**, 331–346.
- McPeck, M. S. (1996). An introduction to recombination and linkage analysis. In *Genetic Mapping and DNA Sequencing*, IMA Volumes in Mathematics and its Applications, vol. 81 (ed. T. Speed & M. S. Waterman), pp. 1–14. New York: Springer-Verlag.
- Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Computer Journal* **7**, 308–313.
- Ott, J. (1991). *Analysis of Human Genetic Linkage*. Baltimore: Johns Hopkins University Press.
- Owen, A. R. G. (1950). The theory of genetical recombination. *Advances in Genetics* **3**, 117–157.
- Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1989). *Numerical Recipes (FORTRAN Version)*. Cambridge: Cambridge University Press.
- Ridout, M. S., Tong, S. & Vowden, C. J. (1997). LINK<sub>3</sub>EM: a program for three-point genetic linkage analysis. version 1.0. East Malling, UK: HRI.
- Ritter, E. & Salamini, F. (1996). The calculation of recombination frequencies in crosses of allogamous plant species with applications to linkage mapping. *Genetical Research* **67**, 55–65.
- Speed, T. P. (1996). What is a genetic map function? In *Genetic Mapping and DNA Sequencing*, IMA Volumes in Mathematics and its Applications, vol. 81 (ed. T. Speed & M. S. Waterman), pp. 65–88. New York: Springer-Verlag.
- Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: JOINMAP. *Plant Journal* **3**, 739–744.

- Stam, P. & Van Ooijen, J. W. (1995). JoinMap version 2.0: software for the calculation of genetic linkage maps. Wageningen: CPRO-DLO.
- Thompson, E. A. (1984). Information gain in joint linkage analysis. *IMA Journal of Mathematics Applied in Medicine & Biology* **1**, 31–49.
- Vowden, C. J. & Ridout, M. S. (1994). LINKEM: a program for genetic linkage analysis, version 1.2. East Malling, UK: HRI.
- Vowden, C. J., Ridout, M. S. & Tobutt, K. R. (1995). LINKEM: a program for genetic linkage analysis. *Journal of Heredity* **86**, 249–250.