Genet. Res., Camb. (1997), **70**, *pp.* 237–250. *With 5 figures.* Printed in the United Kingdom

Linkage analysis in a full-sib family of an outbreeding plant species: overview and consequences for applications

C. MALIEPAARD*, J. JANSEN AND J. W. VAN OOIJEN

Centre for Plant Breeding and Reproduction Research (CPRO-DLO), Centre for Biometry Wageningen (CBW), P.O. Box 16, 6700 AA Wageningen, The Netherlands

(*Received 16 June 1997 and in revised form 31 July 1997*)

Summary

Linkage analysis and map construction using molecular markers is far more complicated in full-sib families of outbreeding plant species than in progenies derived from homozygous parents. Markers may vary in the number of segregating alleles. One or both parents may be heterozygous, markers may be dominant or codominant and usually the linkage phases of marker pairs are unknown. Because of these differences, marker pairs provide different amounts of information for the estimation of recombination frequencies and the linkage phases of the markers in the two parents, and usually these have to be estimated simultaneously. In this paper we present a complete overview of all possible configurations of marker pairs segregating in full-sib families. Maximum likelihood estimators for the recombination frequency and LOD score formulas are presented for all cases. Statistical properties of the estimators are studied analytically and by simulation. Specific problems of dominant markers, in particular with respect to the probability of detecting linkage, the probability of obtaining zero estimates, and the ability to distinguish linkage phase combinations, and consequences for mapping studies in outbred progenies are discussed.

1. Introduction

The application of molecular markers has become a major tool in genetic analysis. Genetic maps are available for a large number of plant and animal species and an increasing number of genes is being detected with the aid of these maps. Various types of markers are used: isozyme markers, restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPDs), amplified fragment length polymorphisms (AFLPs), (sequence tagged) microsatellites, etc. Apart from the techniques, these marker types differ in several respects: number of loci which can be detected, degree of polymorphism within and between accessions and dominance characteristics. Usually, the choice of a particular marker type is based upon these aspects, the convenience of application and, not unimportantly, its costs.

Until recently, linkage analysis with molecular markers in plants has been applied mainly to populations derived from the F_1 of a cross between two fully homozygous diploid parents, i.e. BC_1, F_2, RILs and doubled haploids. One of the reasons is that most important agricultural crops are either self-fertilizing species or inbreeding can be carried out without severe inbreeding depression. Another reason is that linkage analysis is more or less straightforward, while introgression of mapped genes can be done simply by repeated backcrossing.

The differences for linkage analysis between a progeny of a cross derived from inbred lines and a full-sib family of an outbreeding species are due to the number of segregating alleles per locus per parent and the linkage phase of the loci. Segregating populations such as BC_1, F_2 or a set of RILs (in this paper BC_1, F_2 or RILs are considered to be derived from fully homozygous parents) are based on two non-identical inbred lines. Hence, all segregating loci will segregate for only two alleles, and all alleles from the same parent are in coupling phase in the F_1. Contrarily, a cross between two non-identical plants of an outbreeder may segregate for up to four alleles per locus, and this may vary between loci, while the linkage phases usually are unknown.

These differences complicate linkage analysis in a full-sib family (in this paper a full-sib family, or FS-

* Corresponding author. Tel: +31 317 477 004. Fax: +31 317 418 094. e-mail: c.a.maliepaard@cpro.dlo.nl.

family, is considered to be the progeny of a cross between two non-inbred plants of an outbreeding diploid species). There are a number of ways to circumvent these complications and enable the genetic analysis in outbreeders. The most straightforward is the so-called double (or two-way) pseudo-testcross, in which linkage analysis is carried out for each parent separately (Grattapaglia & Sederoff, 1994; Hemmat *et al*., 1994; Grattapaglia *et al*., 1995). However, for crosses in which important alleles segregate in both parents, the integration of the individual parental linkage analyses remains problematic. Another method is to create a backcross progeny in order to simplify the segregation, resembling the BC_1 except for linkage phases, which may be unknown. For crop species with a long juvenile period such as tree species, tulip and lily, this is not a practical solution. Also, incompatibility may block this possibility, or otherwise cause severe selection or inbreeding depression in the progeny.

Linkage analysis using molecular markers in crosses with outbreeders is treated in a number of papers (Ritter *et al*., 1990; Arús *et al*., 1994; Ritter & Salamini, 1996). The latter paper presents formulas useful for the estimation of recombination frequencies in nearly all situations. In some cases the formulas represent the actual estimators, whereas in others the formulas are likelihood equations that have to be implemented in numerical maximization methods such as Newton–Raphson. Unfortunately, two particular configurations were not treated in that paper, although with respect to the estimation, one of these is equivalent to another configuration mentioned. In this paper we present, from a genetic perspective, an overview of the whole range of situations of molecular markers in crosses with outbreeding species. Subsequently, we derive an estimator of the recombination frequency by applying an EM-algorithm to an example configuration. We do this without going into technical detail but completely by explaining the derivation in genetic terms, thereby making the EM-algorithm appear very natural. From this example we generalize the derivation to come to a new, general formula for the estimation of the recombination frequency applicable to all configurations. Using a few comprehensive tables we give a complete overview of the explicit or iterative estimators that were obtained by elaboration of the general formula. These can be implemented easily, even in a computer spreadsheet. A procedure for determining the linkage phases of the parents based upon the progeny is presented. In addition, the quality of the information obtained in the various situations is studied, both analytically and by simulation, and translated into consequences for the application of certain types of markers in linkage analysis for outbreeding species. Finally, we present a new and simple approximation to a confidence interval for a recombination frequency estimate that can be applied to all configurations.

2. Characteristics of the segregation of markers in FS-families

In the two diploid parents of an FS-family up to four different alleles may be present at a single locus; the number of alleles may vary over loci. For all molecular marker types the alleles are usually recognized as fragments with distinct molecular weights. In certain cases a marker detects one or more fragments in some genotypes, whereas it fails to detect a fragment in other genotypes. (Remark: In our terminology a marker is related to a locus, rather than to a single molecular fragment.) The allele corresponding to the absence of a fragment can be called a null-allele. Null-alleles in the parents of an FS-family lead to dominance, i.e. two particular genotypes cannot be distinguished by phenotype. The so-called *segregation type* of a locus, e.g. $ab \times cd$, describes the alleles present in the parents of an FS-family and hence the possible progeny genotypes: the two characters left of the ' × ' represent the alleles of the first parent, the two characters on the right represent those of the second; each distinct allele is symbolized by a different character, and a null-allele with a '0'. Obviously, only segregation types where at least one of the parents is heterozygous are considered for linkage analysis.

In linkage analysis essentially one tries to detect recombination events between loci in both parental meioses. This can be done by reconstructing, for each homologue (or haplotype) of every individual in the offspring, which of the two homologues of one parent contributed to its genotype: a recombination event has occurred if an allele at a certain locus is from one homologue of a parent and the allele at the next locus from the other. This reconstruction uses the phenotypes of the offspring, the parental phenotypes and possibly the grandparental phenotypes. In an FS-family it has to, or can, be done for both parents. If four distinct phenotypes are present in the offspring, the haplotypes which formed these phenotypes can be reconstructed completely, i.e. for each parent the contributed haplotype is clear. This is the case for loci with four alleles ($ab \times cd$, one of the four may be '0'). The segregation types with three non-null-alleles and heterozygous in both parents ($ab \times ac$), or two null-alleles plus two other alleles and heterozygous in both parents ($a0 \times b0$), also allow the complete reconstruction. Therefore, these types are equivalent to the four allele type ($ab \times cd$). For loci heterozygous in only one parent (two alleles: $ab \times aa$, $aa \times ab$, 'a' may be '0', e.g. most RAPD markers; three alleles: $bc \times aa$, $aa \times bc$, one of the three may be '0') the reconstruction can be done for one parent only; these three-allele types are equivalent to the respective two-allele segregation types. Of course, the configuration $a0 \times aa$ does not segregate phenotypically and is not considered. For all other situations the reconstruction can be done only partly. Loci with two alleles and heterozygous in both parents ($ab \times ab$) have two

Table 1. *Configuration numbers of all pairwise combinations of segregation types*

	Locus 2						
Locus 1	$ab \times aa$	$aa \times ab$	$ab \times ab$	$ab \times cd$	$a0 \times a0$	$ab \times a0$	$a0 \times ab$
$ab \times aa$	1	*	2	3	4	5	6
$aa \times ab$		(1)	(2)	(3)	(4)	(6)	(5)
$ab \times ab$			7	8	9	10	(10)
$ab \times cd$				11	12	13	(13)
$a0 \times a0$					14	15	(15)
$ab \times a0$						16	17
$a0 \times ab$							(16)

When no number is given the configuration is equivalent to the configuration with the loci exchanged. When the number is given in parentheses, the configuration is equivalent to its reciprocal cross.
* There is no information on recombination available.

possible parental haplotype combinations for the heterozygous offspring '*ab*': the '*a*' allele may have been derived from either parent and the '*b*' from the alternative. There are even more options for the dominant phenotype '*a*–' for the segregation type $a0 \times a0$: the genotype is either '*aa*' or '*a*0', and in the latter case the '*a*' allele may stem from either parent. This is the typical situation of a RAPD fragment present in both parents and segregating with an expected 3:1 ratio. Finally, the situation with three alleles in which the third allele is a null-allele and heterozygous in both parents ($ab \times a0$, $a0 \times ab$), leaves open for the reconstruction two possibilities for the phenotype '*a*–' in the offspring ('*aa*' or '*a*0'), whereas for the phenotype '*ab*' in the offspring the reconstruction is complete. This situation may occur, for instance, when one of the molecular fragments of a three-allelic RFLP marker is too small to be detected.

Summarizing, in an FS-family there are seven essentially distinct segregation types providing recombination information: (1) two alleles, one parent heterozygous ($ab \times aa$), or (2) the other parent heterozygous ($aa \times ab$), (3) two alleles, both parents heterozygous ($ab \times ab$), (4) four alleles ($ab \times cd$), (5) two alleles, of which one is a null-allele, both parents heterozygous ($a0 \times a0$), (6) three alleles, of which one is a null-allele (in one copy), two parents heterozygous, the null-allele in the one parent ($ab \times a0$), or (7) in the other ($a0 \times ab$). The nine basic configurations of Ritter & Salamini (1996) correspond to these seven segregation types, since four of their configurations (A1A0 × A2A0, A1A2 × A3A0, A1A2 × A1A3 and A1A2 × A3A4) all have the same segregation type: $ab \times cd$, while $ab \times aa$ and $aa \times ab$ are considered equivalent, as are $ab \times a0$ and $a0 \times ab$. The seven segregation types lead to a total of 17 different combinations of loci (Tables 1 and 2), some of which have been well studied, such as the BC_1 type of segregation (no. 1) or the F_2 type of segregation with codominant or dominant markers (nos. 7, 9, 14). The exchange of either the loci or the parents leads to an equivalent situation.

A complicating factor in linkage analysis in crosses with outbreeders is that the linkage phase of the markers will often be unknown *a priori*, while knowledge of the phase is required for the detection of the recombination events. The linkage phase defines the configuration of alleles of a pair of heterozygous loci over the homologous chromosomes in a single parent. It has to be stressed that linkage phase is concerned with the allelic configuration, rather than the loci as such. Additionally, coupling of an allele at locus 1 with a certain allele at locus 2 also means repulsion with the other allele at locus 2. An important distinction from the standard segregating populations with inbred lines is that the linkage phases can be different for the two parents. Also the linkage phase can be undefined in one of the parents due to homozygosity, as in a BC_1. Hence, in an FS-family, we end up with the following linkage phase combinations: coupling (*c*) in the first parent (P_1) and undefined in the second parent (P_2), or vice versa, repulsion (*r*) in P_1 and undefined in P_2, or vice versa, coupling in both parents ($c \times c$), repulsion in both parents ($r \times r$), and coupling in P_1 and repulsion in P_2 ($c \times r$), or vice versa ($r \times c$). For example,

$$\frac{a\ \ a}{b\ \ b} \times \frac{a\ \ b}{b\ \ a}$$

depicts the $c \times r$ combination for a pair of markers with segregation type $ab \times ab$. A linkage phase combination has to be deduced from the segregation in the FS-family itself or from the grandparental genotypes, although this is not always possible.

3. Recombination frequency estimators, LOD scores and determination of linkage phases

Mather (1951), Allard (1956) and Weber & Wricke (1994) developed maximum likelihood estimators of

Table 2. *Definition of marker phenotype indicators*

				Phenotype indicator (f)															
Nr[a]	Locus	P_1^b	P_2^b	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	L1:	*ab*	*aa*	*aa*	*aa*	*ab*	*ab*												
	L2:	*ab*	*aa*	*aa*	*ab*	*aa*	*ab*												
2	L1:	*ab*	*aa*	*aa*	*aa*	*aa*	*ab*	*ab*	*ab*										
	L2:	*ab*	*ab*	*aa*	*ab*	*bb*	*aa*	*ab*	*bb*										
3	L1:	*ab*	*aa*	*aa*	*aa*	*aa*	*aa*	*ab*	*ab*	*ab*	*ab*								
	L2:	*ab*	*cd*	*ac*	*ad*	*bc*	*bd*	*ac*	*ad*	*bc*	*bd*								
4	L1:	*ab*	*aa*	*aa*	*aa*	*ab*	*ab*												
	L2:	*a*0	*a*0	*a*–	00	*a*–	00												
5	L1:	*ab*	*aa*	*aa*	*aa*	*aa*	*ab*	*ab*	*ab*										
	L2:	*ab*	*a*0	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0										
6	L1:	*ab*	*aa*	*aa*	*aa*	*aa*	*ab*	*ab*	*ab*										
	L2:	*a*0	*ab*	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0										
7	L1:	*ab*	*ab*	*aa*	*aa*	*aa*	*ab*	*ab*	*ab*	*bb*	*bb*	*bb*							
	L2:	*ab*	*ab*	*aa*	*ab*	*bb*	*aa*	*ab*	*bb*	*aa*	*ab*	*bb*							
8	L1:	*ab*	*ab*	*aa*	*aa*	*aa*	*aa*	*ab*	*ab*	*ab*	*ab*	*bb*	*bb*	*bb*	*bb*				
	L2:	*ab*	*cd*	*ac*	*ad*	*bc*	*bd*	*ac*	*ad*	*bc*	*bd*	*ac*	*ad*	*bc*	*bd*				
9	L1:	*ab*	*ab*	*aa*	*aa*	*ab*	*ab*	*bb*	*bb*										
	L2:	*a*0	*a*0	*a*–	00	*a*–	00	*a*–	00										
10	L1:	*ab*	*ab*	*aa*	*aa*	*aa*	*ab*	*ab*	*ab*	*bb*	*bb*	*bb*							
	L2:	*ab*	*a*0	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0							
11	L1:	*ab*	*cd*	*ac*	*ac*	*ac*	*ac*	*ad*	*ad*	*ad*	*ad*	*bc*	*bc*	*bc*	*bc*	*bc*	*bd*	*bd*	*bd*
	L2:	*ab*	*cd*	*ac*	*ad*	*bc*	*bd*	*ac*	*ad*	*bc*	*bd*	*ac*	*ad*	*bc*	*bd*	*ac*	*ad*	*bc*	*bd*
12	L1:	*ab*	*cd*	*ac*	*ac*	*ad*	*ad*	*bc*	*bc*	*bd*	*bd*								
	L2:	*a*0	*a*0	*a*–	00	*a*–	00	*a*–	00	*a*–	00								
13	L1:	*ab*	*cd*	*ac*	*ac*	*ac*	*ad*	*ad*	*ad*	*bc*	*bc*	*bc*	*bd*	*bd*	*bd*				
	L2:	*ab*	*a*0	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0				
14	L1:	*a*0	*a*0	*a*–	*a*–	00	00												
	L2:	*a*0	*a*0	*a*–	00	*a*–	00												
15	L1:	*a*0	*a*0	*a*–	*a*–	*a*–	00	00	00										
	L2:	*ab*	*a*0	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0										
16	L1:	*ab*	*a*0	*a*–	*a*–	*a*–	*ab*	*ab*	*ab*	*b*0	*b*0	*b*0							
	L2:	*ab*	*a*0	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0							
17	L1:	*ab*	*a*0	*a*–	*a*–	*a*–	*ab*	*ab*	*ab*	*b*0	*b*0	*b*0							
	L2:	*a*0	*ab*	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0	*a*–	*ab*	*b*0							

The (dominant) phenotypes '*a*–' and '*b*–' can be of genotypes '*a*0' or '*aa*' and '*b*0' or '*bb*', respectively. Reciprocal crosses have identical definitions.

[a] Configuration number according to Table 1.

[b] The genotypes of the two parents ($P_1 \times P_2$) at the first (L1) and the second (L2) locus.

the recombination frequency for a variety of genetic situations in BC_1 and F_2 populations. Ritter *et al.* (1990) developed estimators for most of the genetic situations in crosses between heterozygous parents. Arús *et al.* (1994) contributed the solution to two additional situations, Ritter & Salamini (1996) nearly completed the set, and here we add one more estimator (Table 1, no. 17) and mention a new configuration type (Table 1, no. 6), so that now all combinations with molecular markers with two to four alleles (without epistasis) in an FS-family are covered, including segregation in one or both parents, dominance, and all linkage phase configurations.

In order to calculate the recombination frequency, one needs to know the number of recombination events in both parental meioses. If one knew the genotypes of the gametes, these could be counted easily. However, the marker genotypes of the gametes cannot always be deduced from the phenotypes of the individuals in the progeny. For example, for two $ab \times ab$ loci in linkage phase combination $c \times c$, nine marker phenotypes can be observed in the progeny (table 2, no. 7). Marker phenotypes 1 and 9 are based on two non-recombinant gametes, phenotypes 2, 4, 6 and 8 each on a non-recombinant and a recombinant gamete, and phenotypes 3 and 7 on two recombinant gametes. So, for progeny individuals with one of these marker phenotypes, the number of recombinant gametes can be counted as $(n_1+n_9)0+(n_2+n_4+n_6+n_8)1+(n_3+n_7)2$, where n_f is the number of indi-

viduals with marker phenotype f. However, for the double heterozygous phenotype (5), there are two possible haplotype combinations that cannot be distinguished: either two recombinant or two non-recombinant gametes. We do know, however, the expected proportion of these two combinations in terms of the recombination frequency r, $r^2:(1-r)^2$. Suppose we knew r, then we would know the expected numbers for the two combinations. Using these numbers (and the other n_f), one can estimate the recombination frequency. With this new value the expected numbers for the two combinations can be recalculated, which in turn can be used to estimate a subsequent value of r, and so on. This is an iterative procedure, which can be summarized in the following formula:

$$r_{i+1} = \frac{1}{2n}\left((n_1+n_9)0+(n_2+n_4+n_6+n_8)1 + (n_3+n_7)2+n_5\frac{(1-r_i)^2 0+r_i^2 2}{(1-r_i)^2+r_i^2}\right),$$

$$r_{i+1} = \frac{1}{2n}\left(n_2+n_4+n_6+n_8+2(n_3+n_7) + 2n_5\frac{r_i^2}{(1-r_i)^2+r_i^2}\right),$$

where r_i is the value of r after iteration i. Using an initial value for the recombination frequency (e.g. $r_0 = 0{\cdot}25$), this formula can be iterated until a stable value is reached. Though it may not be obvious here, the previous formula is in fact an ML-estimator of r (Dempster *et al.*, 1977; Lander & Green, 1987).

In the following the above procedure will be formalized in a maximum likelihood context to develop a general formula for the estimators of the recombination frequency in all situations in an FS-family of outbreeders. Any given marker pair will segregate into F phenotypes, with n_1 to n_F individuals in the F phenotype classes adding up to a total of n (Table 2). We define p_f as the probability of (diploid) phenotype f; all p_f are functions of the recombination frequency r. Then, the likelihood of the phenotype frequencies in the progeny is:

$L = \binom{n}{n_1 \ldots n_F}\prod_{f=1}^{F} p_f^{n_f}$, so that:

$$\ln(L) = constant + \sum_{f=1}^{F} n_f \ln(p_f).$$

To maximize the log-likelihood for r we need to solve the likelihood equation:

$$\frac{\partial \ln(L)}{\partial r} = \sum_{f=1}^{F}\frac{n_f}{p_f}\frac{\partial p_f}{\partial r} = 0. \quad (1)$$

For configurations 1, 2, 3, 5, 6, 11, 13, 16 and 17 (Table 1) this likelihood equation can be solved readily leading to explicit ML-estimators (Table 3). For configurations 7 $c\times r$, $r\times c$, and 14 explicit ML-estimators can be derived by substituting θ for r^2, $r(1-r)$ or $(1-r)^2$ in the likelihood equation; in the legitimate range of r the maximum for θ will also be the maximum for r (Table 3). For all remaining configurations, however, the likelihood equations turn into finding zeros of higher-order polynomials, which is difficult. A much easier solution can be obtained by employing the EM-algorithm (Dempster *et al.*, 1977). This approach, as used by Lander & Green (1987) for genetic maps in humans, can be used for all configurations.

Underlying a diploid phenotype of a marker pair is a combination of two haplotypes, i.e. the gametes. Often there can be different haplotype combinations that lead to the same diploid phenotype, e.g. think of linkage phase configuration as in the above example, or dominance. Thus, each of the marker phenotype probabilities, p_f, can be defined by the probabilities π_{fh} of the H_f underlying haplotype combinations:

$$p_f = \sum_{h=1}^{H_f}\pi_{fh}.$$

We can substitute this into (1):

$$\frac{\partial \ln(L)}{\partial r} = \sum_{f=1}^{F}\frac{n_f}{p_f}\sum_{h=1}^{H_f}\frac{\partial \pi_{fh}}{\partial r} = \sum_{f=1}^{F} n_f \sum_{h=1}^{H_f}\frac{\pi_{fh}}{p_f}\frac{\partial \ln(\pi_{fh})}{\partial r} = 0. \quad (2)$$

The probability of a haplotype combination is a simple function of the recombination frequency. A haplotype of two loci is either recombinant or non-recombinant. Recombination can only be observed if there is heterozygosity at both loci in a parent. We define the number of parents heterozygous at both loci to be $\gamma \in \{1,2\}$. Thus, a combination of two haplotypes may consist of zero up to γ recombinants. If α_{fh} and β_{fh} are the numbers of recombinant and non-recombinant haplotypes underlying the haplotype combination h_f, respectively, we obtain the constraint $\alpha_{fh}+\beta_{fh} = \gamma$, with $\alpha_{fh}, \beta_{fh} \in \{0,1,2\}$. Accordingly, the probability for a haplotype combination is $r^{\alpha}(1-r)^{\beta}$, multiplied by a constant. For the derivative of $\ln(\pi_{fh})$ to r we obtain

$$\frac{\partial \ln(\pi_{fh})}{\partial r} = \frac{\alpha_{fh}}{r} - \frac{\beta_{fh}}{1-r} = \frac{\alpha_{fh}-\gamma r}{r(1-r)}. \quad (3)$$

Combining (2) and (3) gives

$$\frac{\partial \ln L}{\partial r} = \frac{1}{r(1-r)}\sum_{h=1}^{F} n_f \sum_{h=1}^{H_f}\frac{\pi_{fh}}{p_f}(\alpha_{fh}-\gamma r) = 0. \quad (4)$$

Now, since both p_f and π_{fh} are functions of r, solving this equation is hard, unless we employ the EM-algorithm. Suppose we know all ratios π_{fh}/p_f, i.e. suppose we know the relative proportion of all underlying haplotype combinations for each phenotype (this is the expectation- or E-step), then we can solve (4) (this is the maximization- or M-step):

$$\hat{r} = \frac{1}{\gamma n}\sum_{f=1}^{F} n_f \sum_{h=1}^{H_f}\frac{\alpha_{fh}\pi_{fh}}{p_f}. \quad (5)$$

Table 3. *ML-estimators of the recombination frequency and LOD score formulas*

No.[a]	Phase[b]	Estimator	LOD score
1 (1)	c	$(n_2+n_3)/n$	$(n_1+n_4)\log(2s)+(n_2+n_3)\log(2r)$
	r	exch[c] $(n_2, n_3; n_1, n_4)$	
2 (2)	c	$(n_3+n_4)/(n_1+n_3+n_4+n_6)$	$(n_1+n_6)\log(2s)+(n_3+n_4)\log(2r)$
	r	exch $(n_3, n_4; n_1, n_6)$	
3	c	$(n_3+n_4+n_5+n_6)/n$	$(n_1+n_2+n_7+n_8)\log(2s)+(n_3+n_4+n_5+n_6)\log(2r)$
	r	exch $(n_3, n_4, n_5, n_6; n_1, n_2, n_7, n_8)$	
(3)	c	exch $(n_3, n_6; n_2, n_7)$	
	r	exch $(n_4, n_5; n_1, n_8)$	
4 (4)	c	$[n_1r/(2-r)+n_2+2n_3r/(1+r)]/n$	$n_1\log(2(1+s)/3)+n_2\log(2r)+n_3\log(2(1+r)/3)+n_4\log(2s)$
	r	exch $(n_1, n_2; n_3, n_4)$	
5 (5)	c	$(n_2+n_3+n_4)/n$	$(n_1+n_5+n_6)\log(2s)+(n_2+n_3+n_4)\log(2r)$
	r	exch $(n_2, n_3, n_4; n_5, n_6, n_1)$	
6 (6)	c	$(n_3+n_5)/(n_2+n_3+n_5+n_6)$	$(n_2+n_6)\log(2s)+(n_3+n_5)\log(2r)$
	r	exch $(n_3, n_5; n_2, n_6)$	
7[d]	$c\times c$	$[n_2+n_4+n_6+n_8+2(n_3+n_7)+2n_5r^2/(1-2rs)]/(2n)$	$2(n_1+n_9)\log(2s)+(n_2+n_4+n_6+n_8)\log(4rs)+2(n_3+n_7)\log(2r)+n_5\log(2(1-2rs))$
	$r\times r$	exch $(n_3, n_7; n_1, n_9)$	
	$c\times r$	$\frac{1}{2}-\sqrt{[\frac{1}{4}-(n_1+n_3+n_5+n_7+n_9)/(2n)]}$	$(n_1+n_3+n_5+n_7+n_9)\log(4rs)+(n_2+n_4+n_6+n_8)\log(2(1-2rs))$
	$r\times c$[e]	exch none	
8	$c\times c$	$[n_2+n_3+n_5+n_8+n_{10}+n_{11}+2(n_4+n_9)+2(n_6+n_7)r^2/(1-2rs)]/(2n)$	$2(n_1+n_{12})\log(2s)+(n_2+n_3+n_5+n_8+n_{10}+n_{11})\log(4rs)+2(n_4+n_9)\log(2r)+(n_6+n_7)\log(2(1-2rs))$
	$c\times r$	exch $(n_1, n_3, n_5, n_7, n_9, n_{11}; n_2, n_4, n_6, n_8, n_{10}, n_{12})$	
	$r\times c$	exch $(n_1, n_2, n_5, n_6, n_9, n_{10}; n_3, n_4, n_7, n_8, n_{11}, n_{12})$	
	$r\times r$	exch $(n_1, n_9; n_4, n_{12})$	
9[d]	$c\times c$	$[2n_1r/(1+r)+2n_2+n_3r(1+r)/(1-rs)+n_4+2n_5/(2-r)]/(2n)$	$n_1\log(4(1-r^2)/3)+2n_2\log(2r)+n_3\log(4(1-rs)/3)+n_4\log(4rs)+n_5\log(4(1-s^2)/3)+2n_6\log(2s)$
	$r\times r$	exch $(n_1, n_2; n_5, n_6)$	
	$c\times r$	$[(n_1+n_5)r(1+r)/(1-rs)+n_2+n_6+2n_3(1-s^2)/(1+2rs)+2n_4r^2/(1-2rs)]/(2n)$	$(n_1+n_5)\log(4(1-rs)/3)+(n_2+n_6)\log(4rs)+n_3\log(2(1+2rs)/3)+n_4\log(2(1-2rs))$
	$r\times c$[e]	exch none	
10	$c\times c$	$[(n_1+2n_4)r+n_2+n_6+n_8+2n_3+2n_5r^2/(1-2rs)+n_7(1+r)]/(2n)$	$(n_1+2n_9)\log(2s)+(n_2+n_6+n_8)\log(4rs)+(2n_3+n_7)\log(2r)+n_5\log(2(1-2rs))$
	$c\times r$	exch $(n_2, n_5, n_8; n_3, n_6, n_9)$	
	$r\times c$	exch $(n_1, n_2, n_3, n_5; n_7, n_9, n_8, n_6)$	
	$r\times r$	exch (n_1, n_3, n_7, n_9)	
(10)	$c\times c$	exch none	
	$c\times r$	exch $(n_1, n_2, n_3, n_5; n_7, n_9, n_8, n_6)$	
	$r\times c$	exch $(n_2, n_5, n_8; n_3, n_6, n_9)$	
	$r\times r$	exch $(n_1, n_3; n_7, n_9)$	

11	$c \times c$	$[n_2+n_3+n_5+n_8+n_9+n_{12}+n_{14}+n_{15}+2(n_4+n_7+n_{10}+n_{13})]/(2n)$	$2(n_1+n_6+n_{11}+n_{16})\log(2s)+(n_2+n_3+n_5+n_8+n_9+n_{12}+n_{14}+n_{15})\log(4rs)$ $+2(n_4+n_7+n_{10}+n_{13})\log(2r)$
	$c \times r$	exch (n_1, n_3, n_5, n_7, n_9, n_{11}, n_{13}, n_{15}; n_2, n_4, n_6, n_8, n_{10}, n_{12}, n_{14}, n_{16})	
	$r \times c$	exch (n_1, n_2, n_5, n_6, n_9, n_{10}, n_{13}, n_{14}; n_3, n_4, n_7, n_8, n_{11}, n_{12}, n_{15}, n_{16})	
	$r \times r$	exch (n_1, n_6, n_{10}, n_{13}; n_4, n_7, n_{11}, n_{16})	
12	$c \times c$	$[2n_1r/(1+r)+2n_2+(n_3+n_5)r(1+r)/(1-rs)+n_4+n_6+2n_7/(2-r)]/(2n)$	$n_1\log(4(1-r^2)/3)+2n_2\log(2r)+(n_3+n_5)\log(4(1-rs)/3)+(n_4+n_6)\log(4rs)$ $+n_7\log(4(1-s^2)/3)+2n_8\log(2s)$
	$c \times r$	exch (n_1, n_2, n_5, n_6; n_3, n_4, n_7, n_8)	
	$r \times c$	exch (n_1, n_2, n_3, n_4; n_5, n_6, n_7, n_8)	
	$r \times r$	exch (n_1, n_2; n_7, n_8)	
13	$c \times c$	$[n_2+n_6+n_7+n_9+n_{10}+n_{11}+2(n_3+n_5)]/[n_1+n_4+n_7+n_{10}+2(n_2+n_3+n_5+n_6+n_8$ $+n_9+n_{11}+n_{12})]$	$(n_1+n_4+2(n_8+n_{12}))\log(2s)+(n_2+n_6+n_9+n_{11})\log(4rs)+(2(n_3+n_5)+n_7+n_{10})$ $\log(2r)$
	$c \times r$	exch (n_2, n_5, n_8, n_{11}; n_3, n_6, n_9, n_{12})	
	$r \times c$	exch (n_1, n_2, n_3, n_4, n_5, n_6; n_7, n_8, n_9, n_{10}, n_{11}, n_{12})	
	$r \times r$	exch (n_1, n_3, n_4, n_5; n_{10}, n_{12}, n_7, n_8)	
(13)	$c \times c$	exch (n_4, n_5; n_7, n_8)	
	$c \times r$	exch (n_1, n_2, n_3, n_5, n_9; n_{10}, n_8, n_6, n_{11}, n_{12})	
	$r \times c$	exch (n_2, n_4, n_5, n_6, n_{11}; n_3, n_7, n_9, n_8, n_{12})	
	$r \times r$	exch (n_1, n_3; n_{10}, n_{12})	
14		$\theta=(n_1-2(n_2+n_3)-n_4)/(2n)+\sqrt{[(n_1-2(n_2+n_3)-n_4)^2/(2n)^2+2n_4/n]}$	$n_1\log(4(2+\theta)/9)+(n_2+n_3)\log(4(1-\theta)/3)+n_4\log(4\theta)$
	$c \times c$	$1-\sqrt{\theta}$ $(\theta=s^2)$	
	$c \times r$	$\frac{1}{2}-\sqrt{(\frac{1}{4}-\theta)}$ $(\theta=rs)$	
	$r \times c^e$	$\frac{1}{2}-\sqrt{(\frac{1}{4}-\theta)}$ $(\theta=rs)$	
	$r \times r$	$\sqrt{\theta}$ $(\theta=r^2)$	
15[d]	$c \times c$	$[n_1r(3-r)/(2-r)+n_2r(1+r)/(1-rs)+2n_3/(2-r)+n_4(1+r)+n_5]/(2n)$	$n_1\log(2(2-r)/3)+n_2\log(4(1-rs)/3)+n_3\log(4(1-s^2)/3)+n_4\log(2r)+n_5\log$ $(4rs)+2n_6\log(2s)$
	$c \times r$	exch (n_2, n_5; n_3, n_6)	
(15)	$c \times c$	exch none	
	$r \times c$	exch (n_2, n_5; n_3, n_6)	
15	$r \times c$	$[n_1r(3+r)/(1+r)+2n_2r/(1+r)+n_3r(1+r)/(1-rs)+n_4r+2n_5+n_6]/(2n)$	$n_1\log(2(1+r)/3)+n_2\log(4(1-r^2)/3)+n_3\log(4(1-rs)/3)+n_4\log(2s)+2n_5\log$ $(2r)+n_6\log(4rs)$
	$r \times r$	exch (n_2, n_5; n_3, n_6)	
(15)	$c \times r$	exch none	
	$r \times r$	exch (n_2, n_5; n_3, n_6)	
16[d]	$c \times c$	$(n_2+n_3+n_4+n_6+n_7+n_8)/[n_1+n_2+n_3+n_4+n_7+2(n_5+n_6+n_8+n_9)]$	$(n_1+2(n_5+n_9))\log(2s)+(n_2+n_3+n_4+n_7)\log(2r)+(n_6+n_8)\log(4rs)$
	$c \times r$	exch (n_5, n_8; n_6, n_9)	
(16)	$c \times c$	exch none	
	$r \times c$	exch (n_5, n_8; n_6, n_9)	
16	$r \times c$	$[n_1+n_5+n_9+2(n_6+n_8)]/[n_1+n_2+n_3+n_4+n_7+2(n_5+n_6+n_8+n_9)]$	$(n_1+2(n_6+n_8))\log(2r)+(n_2+n_3+n_4+n_7)\log(2s)+(n_5+n_9)\log(4rs)$
	$r \times r$	exch (n_5, n_8; n_6, n_9)	
(16)	$c \times r$	exch none	
	$r \times r$	exch (n_5, n_8; n_6, n_9)	
17	$c \times c$	$(n_3+n_6+n_7+n_8+2n_5)/[n_2+n_3+n_4+n_7+2(n_5+n_6+n_8+n_9)]$	$(n_2+n_4+2n_9)\log(2s)+(n_3+n_7+2n_5)\log(2r)+(n_6+n_8)\log(4rs)$
	$c \times r$	exch (n_4, n_5, n_6; n_7, n_8, n_9)	
	$r \times c$	exch (n_2, n_5, n_8; n_3, n_6, n_9)	
	$r \times r$	exch (n_2, n_4, n_5; n_3, n_7, n_9)	

For several linkage phase configurations the estimator and LOD score can be obtained by exchanging as indicated the phenotype frequencies (n_f) in the fully specified preceding formulas. If the estimator contains the recombination frequency itself (as r or s), then it is an iterative estimator. $s=1-r$; n, total number of individuals; log, logarithm to base 10. [a] Configuration number according to Table 1. [b] For each configuration two or four linkage phase combinations are distinguished. [c] exch (n_a, n_b, …; n_w, n_x, …) means: exchange n_a, n_b, … with n_w, n_x, … respectively. [d] The order of the phase combinations and/or reciprocal crosses is changed for convenience. [e] $r \times c$ cannot be distinguished from $c \times r$.

From this we get an estimate of r, and subsequently we can adjust the expectations of the haplotype proportions π_{fh}/p_f and get a new estimate of r, and so on. This iterative procedure is an EM-algorithm (Dempster *et al.*, 1977). In the E-step the ratio π_{fh}/p_f is based upon the value of the recombination frequency of the last iteration, r_i, while in the first iteration usually 0·25 is a good starting value. Table 3 presents this iterative ML-estimator elaborated for the configurations for which an explicit ML-estimator could not be found. For several configurations all phenotypes have just a single underlying haplotype combination, i.e. $H_f = 1$ for all f, so that always $\pi_{fh} = p_f$, and thus (5) becomes an explicit estimator. These situations are special cases of (5) and result in estimators identical to the direct solutions of (1). The use of the EM-algorithm can be extended easily to other more complex situations sometimes encountered in practice, such as where a marker is scored as dominant in part of the progeny and as codominant in the remainder; here, the number of phenotype classes in (5) is simply extended.

To test whether a pair of markers is linked, i.e. $r < 0{\cdot}5$, the LOD score can be used as a test statistic. The LOD score is the logarithm to base 10 of the ratio of the likelihood under the estimated recombination frequency ($r = \hat{r}$) and the likelihood under the null hypothesis of unlinked loci ($r = 0{\cdot}5$): LOD = $\log_{10}(L(r = \hat{r})/L(r = 0{\cdot}5))$. A LOD of 3·0 is commonly used as the threshold for linkage (Morton, 1955; Risch, 1992). Table 3 lists the LOD score formulas for the different configurations.

The use of the estimators of Table 3 presumes that the linkage phase combination is known. However, unlike in crosses with inbred lines, this may not be the case in an FS-family. If the linkage phase combination cannot be determined from the grandparents, then the procedure is to apply the estimators for all possible linkage phases and subsequently deduce the actual linkage phase combination. The method and its success vary for the different configurations. The method depends on (a) the heterozygosity at both loci in both parents, (b) whether both loci have symmetric segregation types ($ab \times ab$, $a0 \times a0$), and (c) whether both loci have a null-allele in the same parent. If the linkage phase combination of a pair cannot be (fully) determined based on this method, then the remaining option is to determine the phases indirectly through combinations with neighbouring loci with more informative segregation types.

Let us first consider the situation where only one of the parents is heterozygous for both markers (configurations 1 to 6). Always $\hat{r}_r = 1 - \hat{r}_c$, with $\hat{r}_r$ the estimate under repulsion and $\hat{r}_c$ the estimate under coupling phase. Of course, only the estimate smaller than 0·5 is a legitimate value. If the LOD score is significant, the linkage phase with the legitimate estimate is chosen.

Next, consider the situations where both loci have a symmetric segregation type (configurations 7, 9, 14). Here, the $c \times r$ and the $r \times c$ estimators are identical, so that the choice between $c \times r$ and $r \times c$ cannot be resolved; also $\hat{r}_{c\times c} = 1 - \hat{r}_{r\times r}$. If the phases are $c \times c$ or $r \times r$ for configurations 7 and 9, then the $c \times r$ (and $r \times c$) estimate is either imaginary (configuration 7) or about 0·5 (configuration 9) with a very low LOD score, while the $c \times c$ or $r \times r$ estimate, respectively, is legitimate. If the phases are $c \times r$ or $r \times c$, then the $c \times c$ or $r \times r$ estimates are about 0·5 with a very low LOD score. Hence, for configurations 7 and 9 the phase combinations $c \times c$ and $r \times r$ can be distinguished from each other and from $c \times r$ or $r \times c$. Configuration 14 is worse, because in addition to being symmetrical, both loci have a null-allele in both parents. Here, if the phases are $c \times c$, then the $c \times r$ (and $r \times c$) estimate is imaginary while the $r \times r$ estimate is larger than 0·5. If, however, the phases are $c \times r$, $r \times c$ or $r \times r$, then all except the $c \times c$ estimate will be legitimate, with identical LOD scores as can be seen from Table 3. Hence, for configuration 14 only the phase combination $c \times c$ on the one hand can be distinguished from $c \times r$, $r \times c$ and $r \times r$ on the other, so that other linked markers, with more informative segregation types, are required to resolve the linkage phase combination.

Subsequently, we examine the non-symmetrical situations where both loci have a null-allele in the same parent (configurations 15, 16). Here, always $\hat{r}_{c\times c} = 1 - \hat{r}_{r\times r}$ and $\hat{r}_{c\times r} = 1 - \hat{r}_{r\times c}$. If the loci are in coupling in the first parent ($c \times c$, $c \times r$), then the estimate for the correct phases has the smallest value and by far the highest LOD score, while the other two estimates are larger than 0·5. If, however, the loci are in repulsion in the first parent ($r \times c$, $r \times r$), then the $r \times c$ and $r \times r$ estimates are approximately equal with similar LOD scores, whilst the other two estimates are larger than 0·5. Hence, the phase combinations $c \times c$ and $c \times r$ can be distinguished from each other and from $r \times c$ or $r \times r$. Although simulations of $r \times c$ and $r \times r$ phases (of configurations 15 and 16) showed that in more than 95% of the significant cases the correct phase combination was estimated, it would be prudent to verify the linkage phases through neighbouring loci (data not shown). This particular behaviour is caused by the typical characteristic of segregation type $ab \times a0$. For the first parent the haplotype contributed to any phenotype in the offspring is always perfectly clear: 'a' or 'b'. For the second parent this depends on the allele contributed by the first parent: if it is 'b' then it is clear, but if it is 'a' then it cannot be resolved whether the second parent contributed the allele 'a' or '0'. Now, suppose two closely linked loci have the segregation type $ab \times a0$ (configuration 16). When they are in coupling in the first parent, nearly half the gametes will have a 'b' allele on both loci, and thus the contribution of the second parent can be determined. For the rest of the gametes of the first parent there will be an 'a' allele at one or both of the

loci, thus blocking the determination of the contribution of the second parent. When, however, the loci are in repulsion in the first parent, then most gametes will have an '*a*' allele at least at one locus, so that the contribution of the second parent cannot be determined for the majority of the phenotypes in the offspring. As a consequence, the phase determination is based on only a small minority of the offspring.

The remaining configurations (configurations 8, 10, 11, 12, 13, 17) supply sufficient information to resolve the linkage phase combination unambiguously. Here, always $\hat{r}_{c\times c} = 1 - \hat{r}_{r\times r}$ and $\hat{r}_{c\times r} = 1 - \hat{r}_{r\times c}$, leaving two legitimate estimates. The estimate with the correct phase practically always has the undoubtedly smaller value and higher LOD, whereas the other legitimate estimate is either close to 0·5 (configurations 8, 11, 12, 17) or in between the smaller estimate and 0·5 (configurations 10, 13).

4. Properties of the ML recombination frequency estimators

In the design of linkage experiments it is important to know the various statistical properties of the recombination frequency estimators for all situations. Bias and variance are important characteristics describing how close one can get to the true value. Another aspect is that segregation types differ in power with respect to detecting linkage; to obtain a complete linkage map it is necessary that linkage is detected for a sufficiently large number of markers at some significance level. Still, when linkage is detected between a pair of loci, this does not necessarily imply that the estimate is accurate. In some marker type combinations significant estimates are predominantly zero estimates, despite the presence of large numbers of recombination events.

In the simulation studies, individuals segregating for two loci were generated according to Mendelian inheritance at a given recombination frequency. Each study was based on 20 000 replicates of F_1 populations consisting of 50, 100, 150, 200 or 1000 individuals. The simulated recombination frequencies ranged from 0 to 0·5 with intervals of 0·001, 0·002 or 0·01. In each F_1 the recombination frequency and the LOD score were calculated using the formulas from Table 3 with the appropriate linkage phase.

(i) *Bias*

For infinite population sizes the ML-estimators of all configurations are unbiased. This was proven analytically for some estimators; for others it was demonstrated by simulation, assuming the linkage phase combination was known, for populations of practically infinite sizes (not shown). However, in practice one deals with finite, sometimes small, population sizes. Here, linkage has to be tested and only recombination frequency estimates with a significant test statistic (the LOD score) are retained. In general, large estimates have small test statistics that are not significant, and as a consequence these large estimates are ignored. Thus, in finite populations, a downward bias is introduced in the set of estimates with a significant LOD score. This is illustrated for a population of 50 individuals in Fig. 1. Since the bias is caused by rejecting non-significant values, it is related to the variance of the estimators, which in turn depends largely on the configuration of the loci as well as on the population size (the variance is treated in the next section). In particular, some of the configurations involving $a0 \times a0$ loci are severely biased due to

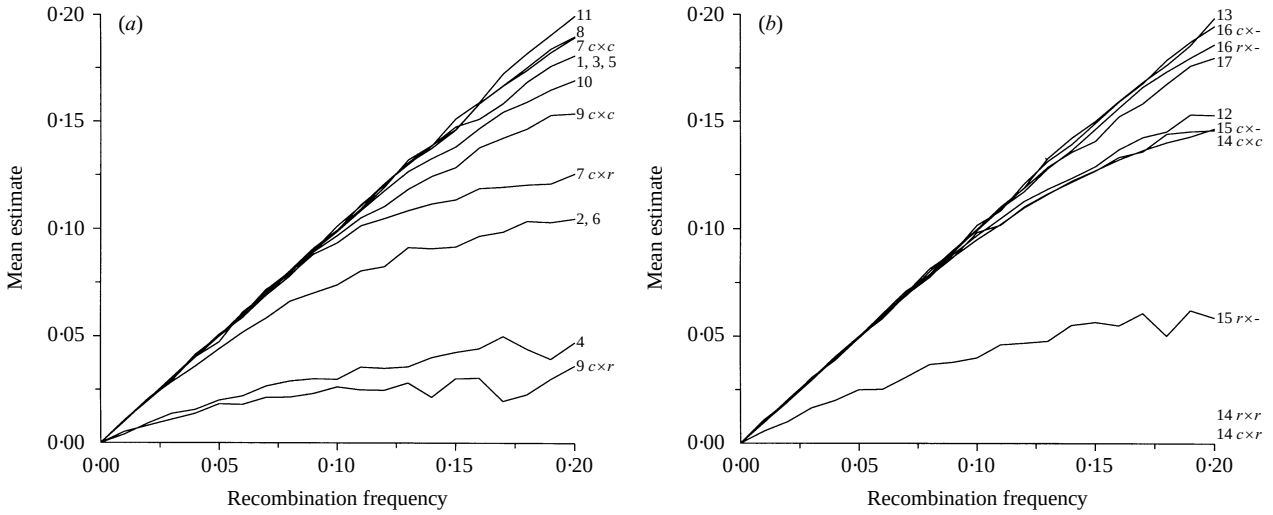

Fig. 1. Means of recombination frequency estimates with a significant LOD (> 3·0) over 1000 simulation runs for each value of the true recombination frequency (steps of 0·01) for a population of 50 individuals. For configurations 7 and 9 the linkage phase combinations $c \times c$ and $c \times r$ are equivalent to $r \times r$ and $r \times c$, respectively, while for configurations 15 and 16 the combinations $c \times c$ and $r \times c$ are equivalent to $c \times r$ and $r \times r$, respectively (indicated with $c \times -$ and $r \times -$). The graphs for configurations 14 $c \times r$ and 14 $r \times r$ coincide with the horizontal axis.

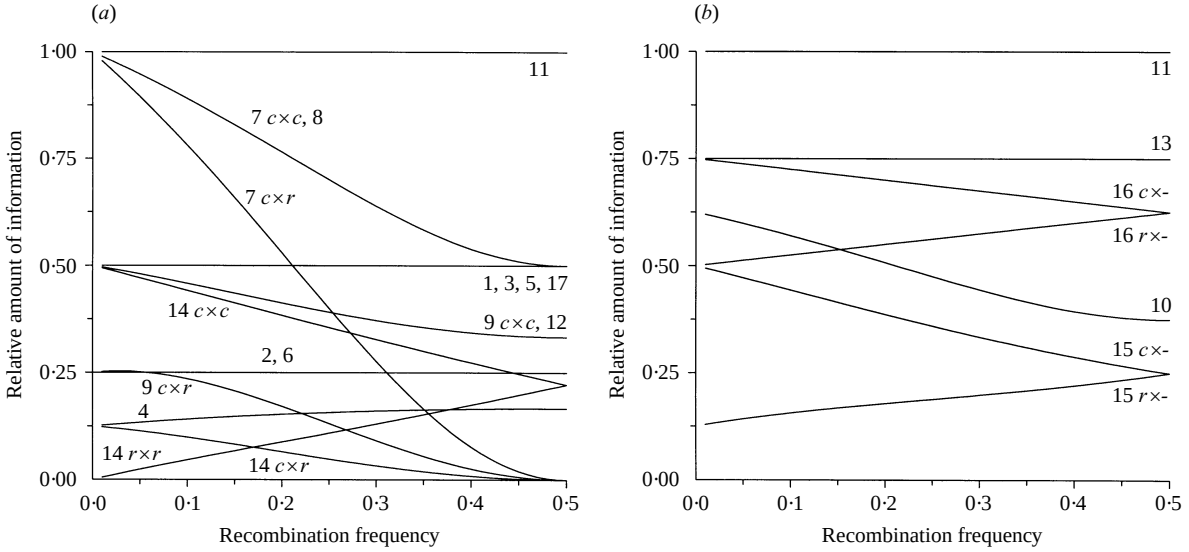

Fig. 2. Information functions relative to configuration 11 for all possible marker configurations in a full-sib family of outbred parents. For configurations 7 and 9 the functions for linkage phase combinations $c \times c$ and $c \times r$ are equal to $r \times r$ and $r \times c$, respectively; for configuration 14 the functions for $c \times r$ and $r \times c$ are equal; for configurations 15 and 16 the combinations $c \times c$ and $r \times r$ are equal to $c \times r$ and $r \times c$, respectively (indicated with $c \times -$ and $r \times -$).

applying the LOD score significance threshold, even with population sizes of 100 or more individuals.

(ii) *Variance*

The variance of a recombination frequency estimator comprises two components: (a) the number of recombination events that created the progeny sample, and (b) the (in)ability with which these events can be detected for a certain configuration of two loci. The first component is determined by the recombination frequency itself and the progeny size; the second by the segregation types of the loci and the linkage phases in the parents. For instance, from a pair of $ab \times cd$ loci all recombination events can be observed perfectly (apart from multiple recombination events); here the variance consists only of the sampling variance. In contrast, from a pair of $a0 \times a0$ loci most of these events cannot be observed directly, but have to be estimated assuming Mendelian ratios. If each $ab \times cd$ locus were completely linked to an $a0 \times a0$ locus, then the estimate based on the two $a0 \times a0$ loci would be different from the estimate using the $ab \times cd$ loci in the same progeny sample.

The variance of ML-estimators is approximately equal to the inverse of Fisher's information, i.e. the expectation of minus the second derivative of the log-likelihood function. Several authors present the information functions of various configurations (Mather, 1951; Allard, 1956; Ritter *et al.*, 1990; Weber & Wricke, 1994; Ritter & Salamini, 1996). The functions relative to the information of configuration 11 (two $ab \times cd$ loci) are depicted in Fig. 2. The information functions of configurations 6 and 17, not described previously, are equal to those of 2 and 1, respectively (equivalent to MCDs 9 and 1 in Ritter & Salamini, 1996). Fig. 2 shows that the combinations with $a0 \times a0$ markers, especially configurations 14 $c \times r$ and $r \times r$, provide a small amount of information. For configurations 2 and 6 (which are equivalent and have the same ML-estimator after exchanging the corresponding phenotype frequencies), the reason for the relatively small amount of information is not so evident. In these configurations, according to expectation half the progeny is not informative at all: the probabilities of two marker phenotype classes are independent of the recombination frequency ($p = \frac{1}{4}$ each) (Ritter & Salamini, 1996). Configuration 4, which is the dominant marker version of configuration 2, is even less informative: here, the non-informative half of the progeny is hidden behind the marker phenotype 'a–' of the $a0 \times a0$ marker and as such increases the variance of the recombination frequency estimate. In configurations 10 and 17 an expected quarter of the progeny is not informative with respect to the recombination frequency.

Since the inverse of Fisher's information is used only as an approximation for the variance, the variance was also investigated by simulation, assuming the linkage phase combination was known and not applying a LOD score threshold. In most instances the approximation was accurate. However, for configuration 14 $r \times r$, the variance estimated from the simulation results was smaller than the inverse information for small values of the recombination frequency. Only for the largest population size tested ($n = 1000$, Fig. 3) did the results agree well with the estimate from the inverse information function. For r approaching 0, the variance estimated from the inverse information function approaches $1/n$. The discrepancy between calculation of the variance from Fisher's information and the simulation results is not well

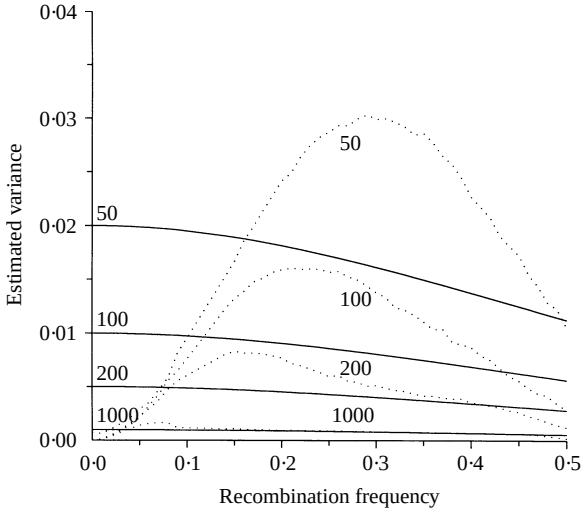

Fig. 3. Variance estimates for the recombination frequency in configuration 14 $r \times r$. Population sizes are 50, 100, 200 and 1000. The continuous lines indicate the variance estimated from the inverse information function; the dotted lines show the variance estimated from recombination frequency estimates over 20000 simulation runs for each value of the true recombination frequency (steps of 0·01).

understood; presumably, this is due to the method being an approximation.

(iii) *Detection of linkage and recombination frequency estimation*

In the development of a linkage map one usually starts with a random set of markers for which no map positions are available. The first step in map construction is the separation of markers into linkage groups. A marker pair is considered to be linked when the marker frequencies obtained in the progeny are significantly different from the expected frequencies in the absence of linkage ($r = 0{\cdot}5$). Several statistics can be used to test linkage, such as Mather's linkage test χ^2_L (Mather, 1951), the contingency test for independence, or the LOD score test. The contingency test for independence is to be recommended, because the other two tests are affected by systematic segregation distortion (Garcia-Dorado & Gallego, 1992). The LOD score test is adequate when there is no systematic segregation distortion, and at present it is possibly the most frequently used test. According to ML-theory, the LOD score follows approximately a chi-square distribution with one degree of freedom: LOD $\sim 0{\cdot}5$ $\log_{10}(e)\,\chi^2_{(1)} = 0{\cdot}22\,\chi^2_{(1)}$ (cf. McCullagh & Nelder, 1989). Often the value 3 is used as the significance threshold, meaning linkage is 1000 times more likely than independent segregation. As a chi-square test, this value corresponds to a significance of 0·0002. This high stringency is needed because many pairs of markers are usually tested (cf. Morton, 1955; Risch, 1992). In the following example we illustrate some important phenomena related to the problems of detecting linkage and the estimation of recombination frequencies.

Suppose we want to construct a map based on RAPD markers determined in an FS-family. These markers would segregate as $ab \times aa$, $aa \times ab$ or $a0 \times a0$, while pairs of markers would be of configurations 1, 4 or 14 in all possible linkage phase combinations (Table 1). Fig. 4 shows that there are large differences between these configurations for the power of detecting linkage. These differences are related to differences in information functions (Fig. 2). For configuration 1, the detection of linkage would usually be no problem, even for recombination frequencies up to 0·3 at a population size of 100. This also holds for configuration 14 $c \times c$ with recombination frequencies up to 0·2. For configuration 4 at a population size of 100, however, the probability of obtaining a significant LOD is larger than 0·9 only for recombination frequencies smaller than 0·1, and the power rapidly decreases beyond 0·1. The power is rather small for configuration 14 $r \times r$, and even dramatically small for configurations 14 $c \times r$ and $r \times c$. Since linkage between $ab \times aa$ and $aa \times ab$ markers cannot be established directly, their linkage has to be determined through $a0 \times a0$ markers, i.e. through configuration 4. Thus, in order to establish linkage between $ab \times aa$ and $aa \times ab$ markers one needs an $a0 \times a0$ marker closely linked to both an $ab \times aa$ and an $aa \times ab$ type marker and hence a large number of $a0 \times a0$ markers would be required; in practice these are not always available.

When significant LOD scores were obtained in our simulations for configurations 14 $r \times r$ and $c \times r$ (and $r \times c$), very often the corresponding estimate of the recombination frequency was zero, which can be understood from the small probability of finding visible recombinants in these configurations. Zero estimates were obtained for even quite large values of the recombination frequency. For instance, for a population size of 150 and a recombination frequency of 0·15, the fractions of the simulation runs that had a significant LOD score were 0·82 and 0·05 for $r \times r$ and $c \times r$, respectively, and the recombination frequency estimate was zero in 51% and 13% of those fractions, respectively. In a population size of 100 the fractions with a significant LOD were 0·45 and 0·02 and zero estimates were found in 87 and 75% of those fractions. In a population of size 50 significant LOD scores were hardly ever found and for zero estimates only.

A more remarkable though very rare phenomenon was the occurrence of non-zero estimates when the true recombination frequency was zero. This was observed in simulations of configurations 14 $r \times r$ and $c \times r$, 4, 9 $c \times r$ and 15 $r \times c$ and $r \times r$. In all cases the frequency of occurrence was below 2% for a population of size 50, and lower for larger populations. This occurred only for situations where there were

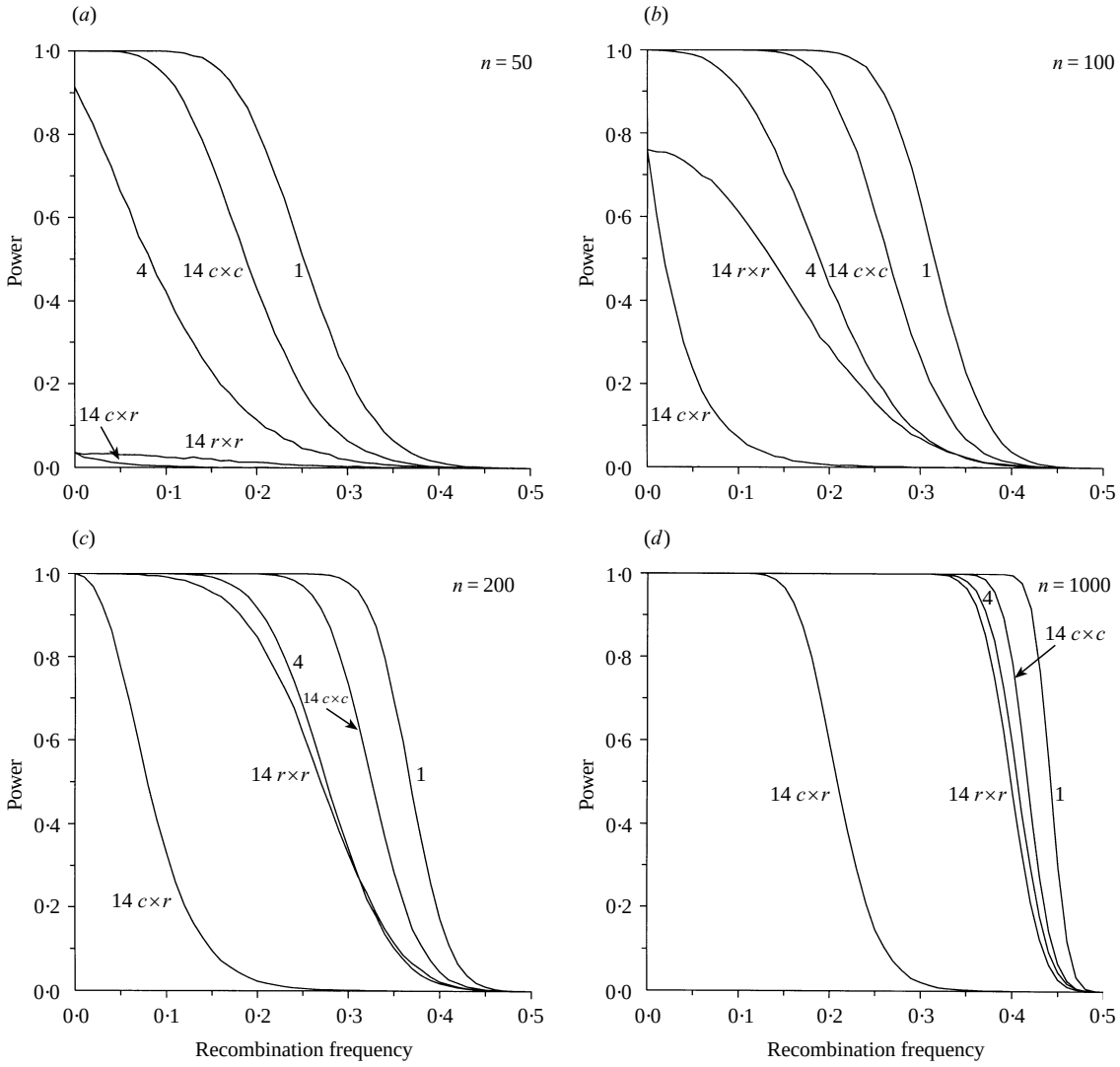

Fig. 4. The power of detecting linkage in configurations involving combinations of $ab \times aa$ (or $aa \times ab$) and $a0 \times a0$ type markers. (Fraction of 20000 simulation runs where a LOD > 3·0 was obtained.) (*a*) Population size $n = 50$; (*b*) $n = 100$; (*c*) $n = 200$; (*d*) $n = 1000$.

large deviations from the expected segregation ratios. It can be proved that this cannot occur for configuration 14 $c \times c$.

Another aspect in this example is the accuracy of the estimates. Although a significant LOD score indicates linkage of a marker pair, it does not imply that the estimate of the recombination frequency is accurate. In the process of mapping we are not only interested in detecting linkage, but accurate estimates are needed to determine the order and distances of the markers. For configuration 1 exact confidence intervals can be given for the recombination frequency, since the number of recombinant genotypes in the progeny follows a binomial distribution with probability r for recombination (Fig. 5). For an estimate of 0·10 and a population size of 100, the 95% confidence interval is [0·05, 0·18]. Although in the other configurations multinominal distributions might be used to construct exact confidence intervals for the recombination frequency, this is quite laborious and these would have to be calculated for each situation separately. Instead, an indication of the accuracy can be obtained by using the relative amount of information from Fig. 2 to construct rough confidence intervals. For instance, for an estimate of 0·10 for configuration 4, the amount of information is a fraction $0{\cdot}13/0{\cdot}50 = 0{\cdot}26$ of the information in configuration 1 at $r = 0{\cdot}10$. So, the 'effective population size' is a fraction 0·26 of the population size for markers in configuration 1. An approximate 95% confidence interval can now be found for a population of size 26 and is equal to [0·02, 0·28]. For configuration 14 $c \times c$ an effective population size of 88 leads to an approximate confidence interval of [0·04, 0·19]. Similarly, approximate confidence intervals of [0·02, 0·32] and [0·01, 0·44] are found for 14 $c \times r$ and $r \times r$, respectively. The width of these rough confidence intervals indicates clearly that difficulties may be expected in the ordering of dominant markers. Although $a0 \times a0$ markers can be used to combine the

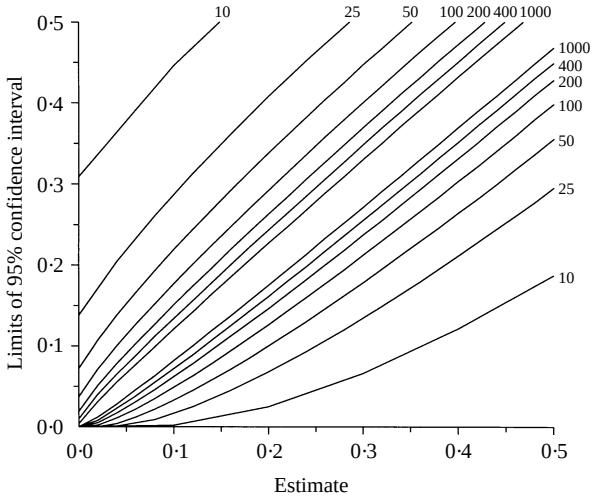

Fig. 5. Lower and upper limits for exact two-sided 95% confidence intervals of the recombination frequency for estimates based on progeny sizes of 10, 25, …, 1000 for configuration 1.

$ab \times aa$ with the $aa \times ab$ markers, their usefulness in establishing the correct marker order between these two groups will be very limited in small populations.

(iv) *Linkage phase*

Prior to the detection of linkage, the linkage phase combination has, of course, to be determined. The success of the methods described previously was tested by simulation. The choice for the linkage phase combination corresponding to a significant LOD score and a legitimate estimate of r was correct in virtually all simulations for all configurations with normal population sizes ($n > 50$), except where linkage phases cannot be distinguished according to theory. In just a few cases indirect estimation (or verification) of the linkage phase through more informative linked markers may be necessary. Of course, if the LOD score is not significant, the choice of the linkage phase cannot be made reliably. From a theoretical point of view it may be interesting to develop a procedure for simultaneous estimation of recombination frequencies and linkage phase combinations over all linked markers. However, in most practical situations this will be redundant.

5. Concluding remarks

This paper describes the marker configurations found in segregating full-sib families of crosses of outbred parents. Seven distinct segregation types characterize the inheritance of individual markers, In practice, the determination of the segregation type of a marker is not always straightforward. For molecular markers this essentially means defining which molecular fragments are allelic. Two marker fragments present in only one parent can be regarded as alleles if either the one or the other is present in all progeny individuals. The probability that this occurs for unlinked loci is very small ($\frac{1}{2}^n$), even in small populations, and also for linked loci this probability (r^n for coupling phase or $(1-r)^n$ for repulsion) decreases rapidly for increasing values of the recombination frequency r. For fragments from different parents, the inference of allelism cannot be made if the markers are based on short DNA sequences and yield large numbers of fragments, such as RAPDs. For RFLPs or sequence tagged sites, however, the inference of allelism will generally be easy since these techniques are based on homology of large segments of DNA and usually yield only a small number of molecular fragments. Still, it has to be realized that in allopolyploid species and species with a polyploid origin, homeology across the genome may impede such conclusions. On the other hand, our own experience on the apple has also shown that for RFLPs the comparison of restriction patterns with different restriction enzymes can be helpful (A. W. van Heusden, CPRO-DLO, Wageningen, the Netherlands, personal communication). If the parental phenotypes are missing and segregation may be distorted, the determination of the segregation type can become complex, e.g. segregation types $a0 \times 00$, $00 \times a0$ and $a0 \times a0$ cannot be distinguished in the progeny since only presence or absence of the band can be scored in the progeny.

In this paper we demonstrated that the various marker pair configurations differ greatly in the accuracy of recombination frequency estimation, the power of detecting linkage and the (im)possibility of estimating the linkage phases in both parents. The information functions as presented in Fig. 2 are a good indication of such differences and may help in the planning of linkage experiments. Also, after collecting marker data the differences in accuracy of the recombination frequency estimates in the various configurations should be considered, so that the ordering of markers per linkage group and the calculation of marker distances may be optimized. After markers have been assigned to linkage groups, conflicting information with respect to the marker order is often provided by the different pairwise recombination frequency estimates. This can be due to missing marker data, but also to random estimation errors in the recombination frequency inherent in the marker configurations. The knowledge of the (in)accuracy of the recombination frequency estimates should then be taken into account to solve such conflicts. For example, in the determination of the distance B–C in a group of four linked markers A–B–C–D, the combined (and weighted) information of the A–B, A–C, B–C, B–D and C–D estimates may well provide a more accurate distance estimate than the single and direct B–C estimate, especially when, for example, markers A and D are of type $ab \times cd$, while B and C are of type $a0 \times a0$. For instance, in the

computer program JoinMap (Stam, 1993; Stam & Van Ooijen, 1995), this is done by using all pairwise recombination frequencies, weighted with the LOD scores, to estimate simultaneously the marker order and distances.

The (in)accuracy of recombination frequency estimates should further be borne in mind when a map resulting from a single cross is used for indirect selection. The upper bound of the confidence interval of the recombination frequency should give an idea of the maximum probability of breaking the linkage between marker and the gene of interest in the subsequent generations. In this respect it is good to note that apart from estimation errors there may also be genetic differences in the recombination frequency (and in the linkage phase combination) in different crosses, as there may be differences between male and female meioses. (e.g. Van Ooijen *et al*., 1994; Plomion & O'Malley, 1996).

The possibility of constructing a single map for a cross, rather than two separate maps for the parents of the cross, depends upon the availability of allelic bridges (Ritter *et al*., 1990). Although in principle $a0 \times a0$ markers could be used as allelic bridges, they will often provide little information so that RAPDs or AFLPs may be of limited use for combining the parental maps. For example, in the double pseudo-testcross populations of apple (Hemmat *et al*., 1994) and *Eucalyptus* (Grattapaglia & Sederoff, 1994), where mostly dominant markers were used, separate maps for the individual parents in the cross could be constructed but the integration of these parental maps was difficult. When a mapping study is done with the intention of integrating the homologous linkage groups of the respective parents, multi-allelic markers, such as RFLPs or microsatellite markers, are recommended. Grattapaglia & Sederoff (1994) and Ritter & Salamini (1996) emphasized the power of such markers for mapping studies in outbred progenies. An extra advantage of these markers is the high probability that they can be used over a wide range of crosses. Another advantage is that, at least where the $ab \times cd$ type of markers is concerned, differences in recombination between the male and the female parent can be estimated directly, whereas, in for example, F_2 populations from inbred lines the recombination frequency has to be assumed equal in the male and female meioses and reciprocal backcross progenies are needed to detect possible differences. If a sufficient number of $ab \times cd$ markers is used in an FS-family of outbred parents, both options are available: either use the separate maps of both parents, or, if the differences in recombination frequency are not too large, construct an integrated map for the cross.

References

Allard, R. W. (1956). Formulas and tables to facilitate the calculation of recombination values in heredity. *Hilgardia* **24**, 235–278.

Arús, P., Olarte, C., Romero, M. & Vargas, F. (1994). Linkage analysis of ten isozyme genes in F_1 segregating almond progenies. *Journal of the American Society for Horticultural Science* **119**, 339–344.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *Series B* **39**, 1–38.

Garcia-Dorado, A. & Gallego, A. (1992). On the use of the classical tests for detecting linkage. *Journal of Heredity* **83**, 143–146.

Grattapaglia, D. & Sederoff, R. (1994). Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. *Genetics* **137**, 1121–1137.

Grattapaglia, D., Bertolucci, F. L. & Sederoff, R. R. (1995). Genetic mapping of QTLs controlling vegetative propagation in *Eucalyptus grandis* and *E. urophylla* using a pseudo-testcross strategy and RAPD markers. *Theoretical and Applied Genetics* **90**, 933–947.

Hemmat, M., Weeden, N. F., Manganaris, A. G. & Lawson, D. M. (1994). A molecular marker linkage map for apple. *Journal of Heredity* **85**, 4–11.

Lander, E. S. & Green, P. (1987). Construction of multilocus genetic maps in humans. *Proceedings of the National Academy of Sciences of the USA* **84**, 2363–2367.

Mather, K. (1951). *The Measurement of Linkage in Heredity*. London: Methuen.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Monographs on Statistics and Applied Probability 37. London: Chapman & Hall.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7**, 277–318.

Plomion, C. & O'Malley, D. M. (1996). Recombination rate differences for pollen parents and seed parents in *Pinus pinaster*. *Heredity* **77**, 341–350.

Risch, N. (1992). Genetic linkage: interpreting LOD scores. *Science* **255**, 803–804.

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.

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. (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.

Van Ooijen, J. W., Sandbrink, J. M., Vrielink, M., Verkerk, R., Zabel, P. & Lindhout, P. (1994). An RFLP linkage map of *Lyopersicon peruvianum*. *Theoretical and Applied Genetics* **89**, 1007–1013.

Weber, W. E. & Wricke, G. (1994). *Genetic Markers in Plant Breeding*. Advances in plant breeding 16. Berlin: Parey Scientific.