Theor Appl Genet (1993) 86: 269–280

Criteria to optimize designs for detection and estimation of linkage between marker loci from segregating populations containing several families

Sijne van der Beek and Johan A. M. van Arendonk

Department of Animal Breeding, Wageningen Agricultural University, P.O. Box 338, NL-6700 AH Wageningen, The Netherlands

Received July 24, 1992; Accepted October 20, 1992
Communicated by L. D. Van Vleck

Summary. Construction of a genome map of highly polymorphic markers has become possible in the past decade. Establishing a complete marker map is an enormous task. Therefore, designs to map molecular markers should be optimal. Designs to detect and estimate linkage between markers from segregating populations were studied. Two measures of design quality were used. The expectation of the maximum lod score indicates the possibility of designs to detect linkage. The accuracy of estimating recombination rate was measured as the probability that the true recombination rate is in a specified internal given the estimate. Accurate approximate methods were developed for rapid evaluation of designs. Seven family types (e.g., double backcross) can be distinguished that describe all families in a segregating population. The family type influences the expected maximum lod score and the accuracy of estimation. The frequency of favorable family types increased with increasing marker polymorphism. At a true recombination rate of 0.20, 27 observations on offspring when five alleles were segregating, and 55 observations on offspring when two alleles were segregating, were necessary to obtain an expected maximum lod score of 3. The probability that the true recombination rate was between 0.15 and 0.25, given an estimate of 0.20, was about 0.85 for a design with 40 families with ten offspring and two alleles segregating and for a design with ten families with ten offspring and six alleles segregating. For smaller designs, accuracies were less, approximate evaluation of accuracy was not justified and, on average, true recombination rates were much greater than estimated given a specified value for the estimated recombination rate.

Key words: Gene mapping – Design – Segregating populations – Detection – Accuracy

Correspondence to: S. van der Beek

Introduction

The construction of a genome map is in progress for several livestock species (e.g., Fries et al. 1989; Bitgood and Somes 1990; Georges et al. 1990; Haley et al. 1990; Brascamp et al. 1991). A map of marker loci, i.e., loci showing Mendelian inheritance, is of use in the further mapping and utilization of loci affecting quantitative traits of economic importance and for the introgression and isolation of genes (Soller and Beckman 1983; Kennedy et al. 1990). Constructing a map of marker loci is laborious, hence optimal experimental designs and efficient statistical procedures are important.

Methods to detect and estimate linkage between loci based on completely inbred lines of plant and animal species have been extensively described (e.g., Mather 1951; Bailey 1961; Ritter et al. 1990). The availability of inbred lines provides a way to optimize the design of experiments to map marker loci. Inbred lines are widely used in plants and laboratory animals. For livestock species completely inbred lines are not available. Methods using information from segregating populations have been developed in human genetics (Morton 1955; Ott 1991). Here, the influence of the researcher on the experimental design is limited. Therefore, emphasis has been on the development of efficient estimation procedures given the data. In livestock species, however, experimental designs can be optimized. In most species many paternal half sibs and full sibs can be obtained in a short period.

In genome mapping experiments a certain set of families is used to map marker loci. In a segregating population a family can be a backcross, an intercross or another type. Once families are selected they are used irrespective of their suitability for a specific pair of markers. Therefore, parameters are needed to measure the effectiveness of detecting and estimating linkage for different family types. The importance of a family type will be determined by its frequency rather than its suitability. An overview of family types, value per family type, frequencies of family types, and other aspects of importance in designing an experiment to map loci in a segregating population, is currently unavailable.

This paper describes factors influencing the quality of designs to map marker loci. Family types will be described systematically. Experimental designs will be compared with respect to the detection of linkage and the accuracy of estimates in segregating populations. Accuracy will be the probability that true recombination rate is in a specified interval given an estimated value for recombination rate. For the detection of linkage an approximate algorithm is developed and evaluated. The accuracies of estimates obtained from two approximate methods are compared with results obtained from simulation. Optimal designs are determined by varying the number and size of full-sib families for different levels of recombination rate and polymorphism of marker loci. In addition, the importance of a knowledge of the linkage phase of marker alleles in the parents is determined.

Notation and assumptions

The linkage relationships between two loci with completely codominant inheritance is studied. Loci are denoted as A and B with alleles $\{A_1, A_2, \ldots\}$ and $\{B_1, B_2, \ldots\}$ respectively. Genotypes are given as $A_1A_1B_1B_2$ when the linkage phase, or simply phase, is unknown and as A_1B_1/A_1B_2 when the phase is known ('/' separates the two haplotypes).

Recombination rate is denoted as θ, its maximum likelihood estimate as $\hat{\theta}$ and the true value as θ_t. The probability of a certain event x is denoted as $P(x)$.

Observations are from full-sib families with information for two generations; i.e., genotypes are known without error for parents and offspring. Full-sib families are assumed to be unrelated and of equal size.

Methods

Estimation of recombination rate

The recombination rate between two markers is estimated by maximum likelihood. The likelihood function for designs with genotype information on parents and offspring of unrelated full-sib families is:

$$\begin{aligned} L(\theta) &= \prod_{f=1}^{N_f} l_f(\theta) = \prod_{f=1}^{N_f} \prod_{p=1}^{n(f)} P(g_{f(p)} | g_{sf}, g_{df}, \theta) \\ &= \prod_{f=1}^{N_f} \sum_{i=1}^{2} \sum_{j=1}^{2} \prod_{p=1}^{n(f)} P(g_{f(p)} | h_{sf}, h_{dfj}, \theta) P(h_{sfi} | g_{sf}) P(h_{dfj} | g_{df}) \end{aligned} \tag{1}$$

where l_f is the likelihood for family f, g_{sf} is the genotype of the sire, g_{df} is the genotype of the dam, $g_{f(p)}$ is the genotype of the p^{th} offspring of family f, h_{sfi} is the haplotype of the sire-given phase i, h_{dfj} is the haplotype of the dam-given phase j, N_f is the number of full-sib families, $n(f)$ is the number of offspring in family f, and θ is the recombination rate.

Three components determine the likelihood function: (1) information on the parental phase; (2) information on the gamete a parent transmits to an offspring; and (3) identification of parental gametes in offspring. These three factors will now be examined in more detail.

(1) For a given genotype two phases are possible each with a probability of 0.5. When the phase for one or both parents is known, e.g., derived from genotypes of grandparents, the likelihood function can be simplified.

(2) An animal with genotype A_iB_k/A_jB_l produces the two non-recombinant gametes A_iB_k and A_jB_l and the two recombinant gametes A_iB_l and A_jB_k. Probabilities are $0.5 \cdot (1-\theta)$ for the two non-recombinant gametes and $0.5 \cdot \theta$ for the two recombinant gametes. When an animal is homozygous for locus A, A_iB_k can not be distinguished from A_jB_k. The probability of a gamete which is either A_iB_k or A_jB_k is $0.5 \cdot (1-\theta) + 0.5 \cdot \theta = 0.5$, i.e., the probability does not depend on θ. The type of a gamete which is either A_iB_k or A_jB_k is unknown and observing such a gamete provides no information about the recombination rate. An animal can produce three types of gametes that are denoted as non-recombinant (non), recombinant (rec) and unknown (un).

(3) Information on the genotype of an offspring and phases in parents is not always sufficient to decide which alleles an offspring inherited from a parent. For example, let A_iB_k/A_jB_l and $A_{i'}B_{k'}/A_{j'}B_{l'}$ be genotypes of two parents and $A_tA_uB_vB_w$ the genotype of an offspring. If $i \neq i'$, $j \neq j'$, $k \neq k'$ and $l \neq l'$ then any gamete the sire produces differs from the gamete the dam produces which enables identification of the parental gametes in offspring. However, if $i = i'$, $j = j'$, $i \neq j$ and $t \neq u$ for locus A then both A_t and A_u could be inherited from either parent and parental gametes cannot be identified.

Offspring can be classified according to the type of gametes received from their parents. Within one class, offspring have equal probability given parental phases, $P(g_{f(p)} | h_{sfi}, h_{dfj}, \theta)$, in equation (1) and the following seven classes can be distinguished:

un, un	two gametes of unknown type
un, non	one gamete of unknown type, one gamete non-recombinant
un, rec	one gamete of unknown type, one gamete recombinant
non, non	two gametes non-recombinant
non, rec	one gamete non-recombinant, one gamete recombinant
rec, rec	two gametes recombinant
2non/2rec	two gametes non-recombinant or two gametes recombinant.

The first class can result from the mating $A_1B_1/A_1B_1 \times A_2B_2/A_2B_2$ where recombinant gametes cannot be distin-

Table 1. Information on gametes inherited and $P(g_{f(p)} | h_{sfi}, h_{dfj}, \theta)$ for the possible genotypes of the offspring of parents with genotypes $A_1B_1/A_2B_2 \cdot A_1B_1/A_1B_2$

Offspring genotype	Type of gametes	$P(g_{f(p)} \| h_{sfi}, h_{dfj}, \theta)$
$A_1A_1B_1B_1 = A_1B_1/A_1B_1$	Un, non	$1/2\,(1-\theta) \cdot 1/2$
$A_1A_2B_2B_2 = A_2B_2/A_1B_2$	Un, non	$1/2\,(1-\theta) \cdot 1/2$
$A_1A_1B_2B_2 = A_1B_2/A_1B_2$	Un, rec	$1/2\,\theta \cdot 1/2$
$A_1A_2B_1B_1 = A_2B_1/A_1B_1$	Un, rec	$1/2\,\theta \cdot 1/2$
$A_1A_1B_1B_2 = A_1B_1/A_1B_2$ or A_1B_2/A_1B_1	Un, un	$1/4\,\theta + 1/4\,(1-\theta) = 1/4$
$A_1A_2B_1B_2 = A_2B_1/A_1B_2$ or A_2B_2/A_1B_1	Un, un	$1/4\,\theta + 1/4\,(1-\theta) = 1/4$

guished from non-recombinant gametes. Such a mating provides no information. Offspring with genotype $A_1A_2B_1B_2$ from the mating $A_1B_1/A_2B_2 \times A_1B_1/A_2B_2$ either inherit two non-recombinant gametes or two recombinant gametes, i.e., class 2non/2rec. Tables 1 and 2 give offspring for two different matings that contain examples for all classes.

For given parental phases two offspring with different genotypes can be in the same class, i.e., have equal probability, $P(g_{f(p)}|h_{sfi}, h_{dfj}, \theta)$, in equation (1). For other parental phases, these two offspring either have equal or different probabilities. Offspring which have equal probabilities independent of parental linkage phase can be grouped in likelihood calculations. All animals within a group have the same contribution to the likelihood function but the contribution of the group might differ between parental phases. Let w_{fk} denote the k^{th} group of offspring of family f. Without loss of information, equation (1) can be written as:

$$L(\theta) = \prod_{f=1}^{N_f} \sum_{i=1}^{2} \sum_{j=1}^{2} \prod_{k=1}^{nw_f} [P(w_{fk}|h_{sfi}, h_{dfj}, \theta)]^{n_{fk}} \cdot P(h_{sfi}|g_{sf})\, P(h_{dfj}|g_{df}) \qquad (2)$$

in which n_{fk} denotes the number of offspring in group w_{fk} and nw_f the number of groups in family f. Expectation of the likelihood can be calculated easier from this equation than from equation (1).

Family types

Families can be divided in seven groups according to possible classes of offspring (Table 3). The probability that a family from a segregating population is of certain type depends on polymorphism of the marker. For instance, any family will be of type I when a marker has only one allele. Assuming Hardy-Weinberg equilibrium and linkage equilibrium, probabilities can be calculated for different family types using the frequencies of the marker alleles.

For most family types the parental phase does not affect the classes to which an offspring can be assigned. For family types IV and V, parents have the same genotypes but differ in phase. The classes to which an offspring can be assigned differ between parental phases for these two family types. When parental phases are unknown, family types IV and V can not be distinguished but have equal probability. All other family types can be distinguished without knowing the parental phases.

Detection of linkage

To determine the strength of evidence in favor of linkage the lod score (Morton 1955) is commonly used. The lod score is defined as:

$$Z(\theta) = \log_{10}\left(\frac{L(\theta)}{L(1/2)}\right). \qquad (3)$$

The maximum value of $Z(\theta)$ is denoted by $Z(\hat{\theta})$.

A maximum lod score of 3 and larger is regarded as significant evidence for linkage. A lod score of 3 approximately equals an 0.05 probability of falsely positive linkage (Morton 1955; Ott 1991). A maximum lod score is not available at the time an experiment is planned. However, the expected maximum lod score can be calculated. The expectation provides a measure of the expected amount of evidence for linkage from a design. The expectation for the maximum lod score, $E[Z(\hat{\theta})]$, given parental genotypes and phases is

$$E[Z(\hat{\theta})] = \sum_{x=1}^{ND} P(D_x)\, Z_x(\hat{\theta}) \qquad (4)$$

Table 2. Information on gametes inherited and $P(g_{f(p)}|h_{sfi}, h_{dfj}, \theta)$ for possible genotypes of offspring of parents with genotypes $A_1B_1/A_2B_2 \cdot A_1B_1/A_2B_2$

Offspring genotype	Type of gametes	$P(g_{f(p)}|h_{sfi}, h_{dfj}, \theta)$
$A_1A_1B_1B_1 = A_1B_1/A_1B_1$	Non, non	$1/2(1-\theta) \cdot 1/2(1-\theta)$
$A_2A_2B_2B_2 = A_2B_2/A_2B_2$	Non, non	$1/2(1-\theta) \cdot 1/2(1-\theta)$
$A_1A_1B_2B_2 = A_1B_2/A_1B_2$	Rec, rec	$1/2\,\theta \cdot 1/2\,\theta$
$A_2A_2B_1B_1 = A_2B_1/A_2B_1$	Rec, rec	$1/2\,\theta \cdot 1/2\,\theta$
$A_1A_2B_1B_1 = A_1B_1/A_2B_1$ or A_2B_1/A_1B_1	Non, rec[a]	$2 \cdot 1/2\,\theta \cdot 1/2(1-\theta)$
$A_1A_2B_2B_2 = A_1B_2/A_2B_2$ or A_2B_2/A_1B_2	Non, rec	$2 \cdot 1/2\,\theta \cdot 1/2(1-\theta)$
$A_1A_1B_1B_2 = A_1B_1/A_1B_2$ or A_1B_2/A_1B_1	Non, rec	$2 \cdot 1/2\,\theta \cdot 1/2(1-\theta)$
$A_2A_2B_1B_2 = A_2B_1/A_2B_2$ or A_2B_2/A_2B_1	Non, rec	$2 \cdot 1/2\,\theta \cdot 1/2(1-\theta)$
$A_1A_2B_1B_2 = A_1B_1/A_2B_2$ or A_2B_2/A_1B_1 or A_1B_2/A_2B_1 or A_2B_1/A_1B_2	2non/2rec	$2 \cdot 1/2(1-\theta)$ $\cdot 1/2(1-\theta) + 2$ $\cdot 1/2\,\theta \cdot 1/2\,\theta$

[a] Both possible combinations of haplotype, given the genotype of the offspring and linkage phases, include a non-recombinant gamete and a recombinant gamete; the factor in front of the probability of this class is due to the two possibilities. For matings with other parental haplotype combinations it is possible that only one haplotype combination including one non and one rec can be derived from the genotype given the haplotypes (eg., $A_1B_1/A_2B_2 \cdot A_3B_3/A_4B_4$ gives $A_1A_3B_2B_4$). $P(g_{f(p)}|h_{sfi}, h_{dfj}, \theta)$ is then $1/2(1-\theta) \cdot 1/2\,\theta$

Table 3. Gamete type inherited by offspring for the seven possible family types

Family type	Type of gametes						
	Un, un	Un non	Un, rec	Non, non	Non, rec	Rec, rec	2 non or 2 rec
I	*						
II	*	*	*				
III		*	*				
IV					*		*
V				*	*	*	*
VI				*	*	*	*
VII				*	*	*	

I – None of the parents is heterozygous for both loci. This family type provides no information about linkage
II – Single backcross, both parents have the same alleles for the intercrossed locus (e.g., $A_1B_1/A_2B_2 \cdot A_1B_1/A_1B_2$
III – Double backcross (e.g., $A_1B_1/A_2B_2 \cdot A_1B_1/A_1B_1$) or single backcross in which the parents have at least one allele not in common for the intercrossed locus (e.g., $A_1B_1/A_2B_2 \cdot A_1B_1/A_1B_3$)
IV – Intercross between parents with the same alleles for both loci and unequal phase (e.g., $A_1B_1/A_2B_2 \cdot A_1B_2/A_2B_1$)
V – Intercross between parents with the same alleles for both loci and equal phase (e.g., $A_1B_1/A_2B_2 \cdot A_1B_1/A_2B_2$)
VI – Intercross between parents with the same alleles for one locus and at most one allele in common for the other locus (e.g., $A_1B_1/A_2B_2 \cdot A_1B_1/A_2B_3$)
VII – Intercross between parents in which both loci have at most one allele in common (e.g., $A_1B_1/A_2B_2 \cdot A_1B_1/A_3B_3$)

in which D_x is the data in realization x, ND is the number of possible realizations of data, $P(D_x)$ is the probability for D_x given θ and phases, and $Z_x(\hat{\theta})$ is the maximum lod score for data set x.

The $E[Z(\hat{\theta})]$ for a design with unknown phases is the weighted average of the $E[Z(\hat{\theta})]$ for all the possible phases.

The number of data sets to be considered in an exact calculation of $E[Z(\hat{\theta})]$ increases rapidly with the number of families and number of offspring per family. Computational requirements soon exceed a practical level. An approximate method to calculate $E[Z(\hat{\theta})]$, based on distributional properties of the lod score, is used.

The expectation of $Z(\hat{\theta})$ over all realizations of data can be written as

$$\begin{aligned} E[Z(\hat{\theta})] &= E\left(\log_{10}\frac{L(\hat{\theta})}{L(\theta_t)}\,\frac{L(\theta_t)}{L(1/2)}\right) \\ &= E\left(\log_{10}\frac{L(\hat{\theta})}{L(\theta_t)}\right) + E\left(\log_{10}\frac{L(\theta_t)}{L(1/2)}\right) \qquad (5) \end{aligned}$$

The term $2\ln\frac{L(\hat{\theta})}{L(\theta_t)}$ has asymptotically a χ^2 distribution (Kendall and Stuart 1978) with an expectation over all data sets of 1. Equivalence between $0.217\cdot 2\ln\frac{L(\hat{\theta})}{L(\theta_t)}$ and $\log_{10}\frac{L(\hat{\theta})}{L(\theta_t)}$ leads to an approximation of the first part of (5):

$$E\left(\log_{10}\frac{L(\hat{\theta})}{L(\theta_t)}\right) = 0.217\,E\left(2\ln\frac{L(\hat{\theta})}{L(\theta_t)}\right) \approx 0.217. \qquad (6)$$

The second part of (5) $\left[E\left(\log_{10}\frac{L(\theta_t)}{L(1/2)}\right)\right]$, is equal to the expected lod score (Ott 1991). The expected lod score is additive over families (Ott 1991) and can be calculated as:

$$\begin{aligned} E\left(\log_{10}\frac{L(\theta_t)}{L(1/2)}\right) &= E\left(\log_{10}\prod_{f=1}^{N_f}\frac{l_f(\theta_t)}{l_f(1/2)}\right) \\ &= \sum_{f=1}^{N_f} E\left(\log_{10}\frac{l_f(\theta_t)}{l_f(1/2)}\right) \qquad (7) \end{aligned}$$

where l_f represents the likelihood for family f as defined by (1).

The fact that families of the same type and with an equal number of offspring have equal expectation can be used for further simplification:

$$\begin{aligned} \sum_{f=1}^{N_f} E\left(\log_{10}\frac{l_f(\theta_t)}{l_f(1/2)}\right) &= \sum_{i=1}^{7} m_i E\left(\log_{10}\frac{l_i(\theta_t)}{l_i(1/2)}\right) \\ &= \sum_{i=1}^{7} m_i \sum_{y=1}^{nd_i} P(d_{iy})\log_{10}\left(\frac{l_i(\theta_t\,|\,d_{iy})}{l_i(1/2\,|\,d_{iy})}\right) \qquad (8) \end{aligned}$$

where i denotes family type, m_i the number of families of type i, d_{iy} the realization y of data for all families of type i, nd_i the number of possible realizations of data for a family of type i, and $P(d_{iy})$ the probability of realization of d_{iy}.

Combining (6) and (8) results in an approximate $E[Z(\hat{\theta})]$, $E[Z(\hat{\theta})_{ap}]$:

$$E[Z(\hat{\theta})_{ap}] = 0.217 + \sum_{i=1}^{7} m_i \sum_{y=1}^{nd_i} P(d_{iy})\log_{10}\left(\frac{l_i(\theta_t\,|\,d_{iy})}{l_i(1/2\,|\,d_{iy})}\right). \qquad (9)$$

In segregating populations, family type of a family is unknown at the start of an experiment and the exact value of m_i is unknown. The expectation for m_i is calculated as the product of the number of families and the probability that a family is of type i.

Exact calculation of the $E[Z(\hat{\theta})]$ involves the maximization of many ($\prod_{f=1}^{NF} nd_f$ where nd_f is the number of realizations of data for family f) likelihood functions. This number is reduced considerably with approximation (9), i.e., $2\sum_{i=1} nd_i$.

For families of type I, II, III, VI or VII, $E[Z(\hat{\theta})]$ is independent of the phases. To calculate $E[Z(\hat{\theta})]$ in these cases an arbitrary phase is assigned to a family when phases are unknown. When a family can be either type IV or V given its parental genotypes, $E[Z(\hat{\theta})]$ is calculated for both cases. The average of both possibilities is used as an approximation to $E[Z(\hat{\theta})]$.

Accuracy of estimation

The accuracy of the estimated recombination rate $(\hat{\theta})$ is measured as the probability that true recombination rate (θ_t) is in a specified interval given the estimate $(\hat{\theta}=x)$. This probability is calculated as:

$$P(y_1<\theta_t<y_2\,|\,\hat{\theta}=x) = \frac{P(\hat{\theta}=x\,|\,y_1<\theta_t<y_2)}{P(\hat{\theta}=x)} \qquad (10)$$

where

$P(\hat{\theta}=x\,|\,y_1<\theta_t<y_2) = \int_{y_1}^{y_2} P(\hat{\theta}=x\,|\,\theta_t)\,f(\theta_t)\,d(\theta_t)$ and

$P(\hat{\theta}=x) = \int_{0}^{0.5} P(\hat{\theta}=x\,|\,\theta_t)\,f(\theta_t)\,d(\theta_t)$ with

y_1 and y_2 = lower and upper limits for θ_t, respectively and $f(\theta_t)$ = prior density function of θ_t.

Calculation of (10) involves the use of the probability density of estimates given the true recombination rate to calculate the probability that an estimate is in a certain range given θ_t and a prior probability density function of θ_t.

The maximum likelihood estimate is asymptotically normally distributed and has asymptotic variance equal to the inverse of the expected information (Kendall and Stuart 1978). Information is defined as the second derivative of the likelihood function. When parental linkage phases are known for a family of given type, expected information is a linear function of the number of offspring in the family. For example, a family of family type III with n offspring has expected information $n/(\theta\cdot(1-\theta))$. The relation between the number of offspring and the expected information is nonlinear when parental phases are unknown.

The probability that an estimate is in a certain range for a given θ_t is approximated assuming a normal distribution with variance equal to the inverse of the expected information. The same approximation is used for the probability that the true recombination rate is in a certain range for a given value of the estimated recombination rate.

The prior density function of the recombination rate between marker loci depends on several factors: the number of chromosomes of the species, the lengths of chromosomes, the physical distribution of marker loci on chromosomes, and the relation of distance and recombination rate between loci.

The following is assumed: loci have a probability of 1/20th to be located on the same chromosome, loci are uniformly distributed over chromosomes with a length of 1 morgan, and map distance and recombination rate are related by Haldane's (1919) mapping function. Following the approach of Morton (1955) the prior density function of θ_t is:

$$\begin{aligned} f(\theta_t) &= 0.05\left(\frac{2}{1-2\theta_t}(0.5\ln(1-2\theta_t)+1)\right) && \text{for } 0\le\theta_t<0.432; \\ f(\theta_t) &= 0 && \text{for } 0.432\le\theta_t<0.5; \\ f(\theta_t) &= 0.95 && \text{for } \theta_t=0.5. \qquad (11) \end{aligned}$$

Simulation

The average value of the maximum lod score, the distribution of the maximum likelihood estimator of θ, and the distribution of θ_t for a given value of $\hat{\theta}$ were all obtained using Monte Carlo simulation.

The average maximum lod score was calculated for all seven family types for different values of θ and a different number of offspring per family. In each simulated data set the number of offspring in each group w_{fk} was simulated using the probabilities of groups. The probability of genotype group w_{fk} for a given family type depends on θ. Values of 0.05 and 0.20 for θ and numbers of offspring of 2, 4, 8, 16 and 32 were used. For each alternative, 1,000 data sets were simulated. In each data set $Z(\hat{\theta})$ was computed and averaged to obtain $E[Z(\hat{\theta})]$.

The distribution of the maximum likelihood estimator of θ was calculated from estimated recombination rates in different replicates. Recombination rate was estimated from data containing information from several families. The probability that a family is of a given type was calculated from the frequencies of marker alleles assuming Hardy-Weinberg equilibrium for individual loci and linkage equilibrium between loci. These probabilities were used to simulate family types for a data set. The following values were used; for θ: 0.05, 0.20 and 0.50; for the number of families: 1, 2, 4, 8, 16 and 32; for the number of offspring per family: 2, 4, 8, 16, 32. For each alternative, 10,000 data sets were simulated.

The distribution of θ_t for a given value of $\hat{\theta}$ was determined for a given number of offspring per family and a given number of families. Simulations to obtain the distribution were as follows. First, two markers were randomly located on a chromosome of 1 morgan using a uniform distribution. The recombination rate between the markers was calculated from the distance between the markers assuming Haldane's (1919) mapping function. Family types and information on offspring were simulated for the simulated θ_t and $\hat{\theta}$ calculated from the data. This was repeated 100,000 times. The number of realizations of $(\hat{\theta}=x, \theta_t=y)$ were counted. Second, 100,000 data sets were simulated with markers located on different chromosomes, i.e., for $\theta_t=0.50$. Maximum likelihood estimates were calculated and the number of realizations of $(\hat{\theta}=x, \theta_t=0.50)$ counted. The number of realizations of $(\hat{\theta}=x, \theta_t=0.50)$ were multiplied by 19 to take into account the prior probability that two markers are on separate chromosomes is 19 times the prior probability that two markers are on the same chromosome. Obtained from counts of $(\hat{\theta}=x, \theta_t=y)$ were: the distribution of θ_t given $\hat{\theta}=x$; the distribution of $\hat{\theta}$ given $\theta_t=y$; $P(y_1<\theta_t<y_2|\hat{\theta}=x)$; the average value of θ_t given $\hat{\theta}=x$ and the average value of $\hat{\theta}$ given $\theta_t=y$. The following values were used in simulations; for number of families: 10, 20 and 40; for number of offspring per family: 4, 10 and 46.

Results

Detection

Figure 1 shows the approximated expected maximum lod score, $E[Z(\hat{\theta})_{ap}]$, and the expected maximum lod score obtained by simulation, $E[Z(\hat{\theta})_{sim}]$, for alternatives with 4–16 families and 4–16 offspring per family. The $E[Z(\hat{\theta})_{ap}]$ agree with the expectation obtained by simulation. The $E[Z(\hat{\theta})_{ap}]$ will be used in this study and called $E[Z(\hat{\theta})]$ in the remainder of this paper.

Probabilities of family types were calculated assuming a population in Hardy-Weinberg equilibrium. The number of alleles per marker locus influenced the distribution of families over family types (Table 4). The probability of family type I (least favorable family type) decreased and the probability of type VII (most favorable family type) increased when the number of alleles increased. The heterozygosity of marker loci increased with an increasing number of alleles. The marginal change per

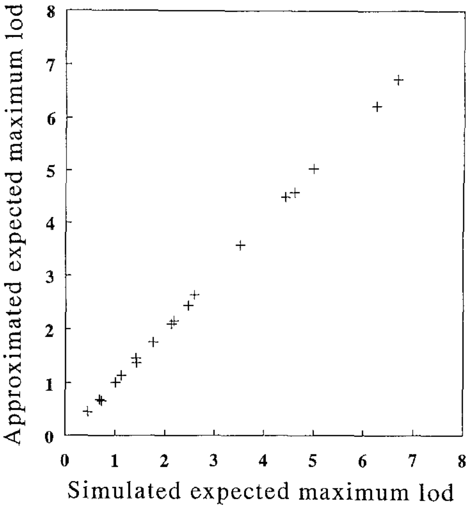

Fig. 1. Relation between simulated expected maximum lod score and approximated expected maximum lod score based on designs with 4–16 unrelated full-sib families with 4–16 offspring where each (+) represents a design

Table 4. Probability of family types for a varying number of equiprobable alleles

No. alleles	% Heterozygous	Type I	Type II	Type III	Type IV	Type V	Type VI	Type VII
2	50	0.56	0.25	0.13	0.03	0.03	0.00	0.00
3	67	0.31	0.09	0.41	0.01	0.01	0.09	0.09
4	75	0.19	0.04	0.46	0.00	0.00	0.09	0.22
5	80	0.13	0.02	0.44	0.00	0.00	0.07	0.33
6	83	0.09	0.01	0.42	0.00	0.00	0.06	0.42
7	86	0.07	0.00	0.39	0.00	0.00	0.05	0.49
8	88	0.05	0.00	0.36	0.00	0.00	0.04	0.55
9	89	0.04	0.00	0.33	0.00	0.00	0.03	0.59
10	90	0.03	0.00	0.31	0.00	0.00	0.03	0.63

Table 5. Expected maximum lod score (E [Z $(\hat{\theta})$]) for designs with one family for different family types and number of offspring per family, for two values of θ_t and when phase is known; E [Z $(\hat{\theta})$] for phase known minus E [Z $(\hat{\theta})$] when phase is unknown is given between brackets

No. offspring	Family type					
	II	III	IV	V	VI	VII
$\theta_t = 0.05$						
4	0.46 {0.19}	1.02 {0.29}	0.87 {0.26}	1.40 {0.54}	1.40 {0.46}	1.93 {0.59}
8	1.01 {0.28}	1.93 {0.30}	1.57 {0.30}	2.62 {0.60}	2.62 {0.60}	3.68 {0.60}
16	1.93 {0.30}	3.68 {0.30}	2.88 {0.30}	5.00 {0.60}	5.00 {0.60}	7.12 {0.60}
32	3.68 {0.30}	7.12 {0.30}	5.50 {0.30}	9.74 {0.60}	9.74 {0.60}	13.98 {0.60}
$\theta_t = 0.20$						
4	0.31 {0.14}	0.59 {0.22}	0.35 {0.14}	0.65 {0.35}	0.65 {0.20}	0.92 {0.41}
8	0.57 {0.19}	0.92 {0.25}	0.46 {0.17}	1.03 {0.46}	1.03 {0.31}	1.57 {0.53}
16	0.92 {0.25}	1.57 {0.28}	0.68 {0.21}	1.80 {0.55}	1.80 {0.43}	2.90 {0.59}
32	1.58 {0.29}	2.90 {0.30}	1.14 {0.26}	3.36 {0.60}	3.36 {0.52}	5.58 {0.60}

additional allele decreased with an increasing number of alleles for both the heterozygosity and the probabilities of family types. The results in Table 4 show a clear relation between the distribution over family types and heterozygosity. This relation is expected to hold when the number of alleles differs per locus and when alleles have unequal frequencies.

Table 5 gives the relation between E [Z $(\hat{\theta})$] and family type, the number of offspring, θ_t, and knowledge of the parental phases. When phases were known, the E [Z $(\hat{\theta})$] of family types II, III and VII were in the proportion of 0.5:1:2 independent of θ_t and the number of offspring. This proportion corresponds to the number of informative gametes for these family types. The ratio between the E [Z $(\hat{\theta})$] for family type IV and family type III was close to 0.8 when the recombination rate was 0.05. The ratio was 0.4 when the recombination rate was 0.20. For family types V and VI, ratios with family type III were 1.4 when the recombination rate was 0.05 and 1.15 when the recombination rate was 0.20. For family types IV, V and VI, there was no direct relation between the proportional E [Z $(\hat{\theta})$] and the number of gametes.

The difference in E [Z $(\hat{\theta})$] due to a knowledge of phase approached to a constant for each family type with increasing family size (Table 5). As an explanation, consider the function for the lod score for one double backcross family (family type III):

$$\mathrm{lod}_{\mathrm{known}}(\theta) = \log_{10}\left(\frac{\theta^x(1-\theta)^{(n-x)}}{0.5^n}\right)$$

$$= x\log_{10}(\theta) + (n-x)\log_{10}(1-\theta) + n\log_{10}(2)$$

$$\mathrm{lod}_{\mathrm{unknown}}(\theta) = \log_{10}\left(\frac{0.5\,(\theta^x(1-\theta)^{n-x} + (1-\theta)^x\theta^{n-x})}{0.5^n}\right)$$

$$= \log_{10}\left(\frac{1}{2}\right) + \log_{10} \cdot \left(\theta^x(1-\theta)^{(n-x)}\left(1+\left(\frac{\theta}{1-\theta}\right)^{(n-2x)}\right)\right) + n\log_{10}(2)$$

$$= \log_{10}\left(\frac{1}{2}\right) + \log_{10}\left(1+\left(\frac{\theta}{1-\theta}\right)^{n-2x}\right) + x\log_{10}\theta + (n-x)\log_{10}(1-\theta) + n\log_{10}(2)$$

$$= \mathrm{lod}_{\mathrm{known}}(\theta) - 0.3 + \log_{10}\left(1+\left(\frac{\theta}{1-\theta}\right)^{n-2x}\right)$$

where x is the number of recombinant gametes, $\mathrm{lod}_{\mathrm{known}}$ is the function for lod score when the phase is known, and $\mathrm{lod}_{\mathrm{unknown}}$ is the function for lod score when the phase is unknown. The expectation for x is $n^*\theta_t$. The term

$$\log_{10}\left(1+\left(\frac{\theta}{1-\theta}\right)^{n-2x}\right)$$

goes to zero when n becomes large and θ_t is not close to 0.5. The difference between the lod score for phase known and phase unknown is then a constant (0.3), which equals to the difference in E [Z $(\hat{\theta})$]. Similar relations occur for families of other types. The difference in E [Z $(\hat{\theta})$] approached 0.3 for family types II, III and IV and 0.6 for family types V, VI and VII.

The additional number of observations on offspring needed to compensate for a smaller E [Z $(\hat{\theta})$] due to lack of knowledge of phases was dependent on family type and θ_t. Therefore, the additional number of observations on offspring needed for an average family depended on polymorphism of the marker and θ_t (Table 6). The maximum of 5.5 occurred when the number of alleles was two and θ_t was 0.2. The alternative to typing additional offspring is typing grandparents (four additional observations). Partial or complete knowledge of phases can be obtained from grandparents.

Table 6. Additional number of observations[a] on offspring to compensate for a smaller expected maximum lod score due to an unknown phase for an average informative[b] family from populations with 2, 5 or 10 equiprobable alleles and a θ_t of 0.05 or 0.20

No. alleles	Known – unknown[c]	θ_t	
		0.05	0.20
2	0.141	1.9	5.5
5	0.336	1.4	3.6
10	0.462	1.4	3.6

[a] Average difference in E[Z($\hat{\theta}$)] divided by the average E[Z($\hat{\theta}$)] per observation on offspring. Average E[Z($\hat{\theta}$)] per observation calculated as (E[Z($\hat{\theta}$)] at 32 offspring – E[Z($\hat{\theta}$)] at 16 offspring)/16

[b] An informative family is not of type I

[c] Average difference in E[Z($\hat{\theta}$)] = difference in E[Z($\hat{\theta}$)] for family type II (=0.3) · probability family type II + difference in E[Z($\hat{\theta}$)] for family type III (=0.3) · probability family type III + etc.

Table 7. Number of observations[a] on offspring for an E[Z($\hat{\theta}$)] of 3 for an average informative family from populations with 2, 5 or 10 alleles of equal frequency and a θ_t of 0.05 or 0.20[b]

No. alleles	θ_t	
	0.05	0.20
2	19	55
5	10	27
10	8	22

[a] Excluding observations on parents

[b] For each family type E[Z($\hat{\theta}$)] per observation is calculated from the difference between E[Z($\hat{\theta}$)] with 32 offspring and E[Z($\hat{\theta}$)] with 16 offspring. A weighted average E[Z($\hat{\theta}$)] per observation is calculated using the probabilities for the family types given the number of alleles

Table 8. Expected maximum lod score dependent on the number of families and the total number of observations[a] on offspring for designs with family types due to chance (both marker loci have two alleles), unknown phases, and $\theta_t = 0.05$

No. observations	Number of families				
	1	2	4	8	16
32	2.26	2.11	1.85	1.43	0.90
64	4.44	4.30	4.01	3.48	2.64

[a] Excluding observations on parents

The average number of offspring in one family needed to obtain an E[Z($\hat{\theta}$)] of 3 is given in Table 7 for three levels of marker polymorphism and two values of θ_t. The required number of observations on offspring was 19 for a θ_t of 0.05 and 55 for a θ_t of 0.20 when the number of alleles was two, which is twice the number required with five alleles.

A given number of observations on offspring can be obtained by analyzing different numbers of families. E[Z($\hat{\theta}$)] decreased if the number of families increased and phase was unknown (Table 8). For each family, information is used to estimate the phase. Therefore, the required number of observations to obtain an E[Z($\hat{\theta}$)] of 3 will increase when observations on offspring are divided over more than one family. Further, the number of observations increases with the number of families because for each family two parents must be genotyped.

Accuracy

In Table 9 the mean and standard error for $\hat{\theta}$ are given for different numbers of observations, different θ_t values, and different full-sib family size. Observed standard errors were obtained from replicated simulations. Standard errors were approximated using the second derivative of the correct likelihood function (σ_{ap2}) and the likelihood function where parental phases were assumed known (σ_{ap1}). Estimated recombination rates were biased upward for a θ_t of 0.2. For a θ_t of 0.05 an upward bias was observed when the number of animals per family was 4. Downward bias was found for a θ_t of 0.5 which is inevitable because estimated recombination rates are restricted to be between 0 and 0.5. The bias diminished with an increasing number of observations. For a given number of observations, bias was less when the number of observations per family increased. Observed and approximated standard errors agree closely for designs with 120 or more observations when θ_t is 0.05 and 240 or more observations when θ_t is 0.20. With fewer observations both approximations underestimated the standard error. Approximate standard errors using the correct likelihood function (σ_{ap2}) were closer to the observed standard errors for a θ_t of 0.2. Expected information, calculated from the correct likelihood function, is zero for unlinked loci and, as a result, σ_{ap2} does not exist for a θ_t of 0.5.

The observed cumulative probability distribution of the estimates is compared with the cumulative normal distribution in Fig. 2. For ten families of four offspring and two alleles for each locus, observed probabilities for estimates of 0 or 0.5 were larger than probabilities calculated from the normal distribution. Differences might be expected since the normal distribution function is only approximate for large numbers because of the central limit theorem. With increasing numbers the deviation between the approximation and the observed distribution became smaller and was negligible for 400 observations. Probabilities that estimates were in a certain interval, given a true recombination rate, could be adequately approximated using the normal distribution for larger designs.

Gene maps or parts of gene maps are often evaluated based on spacing between considered loci. A logical as-

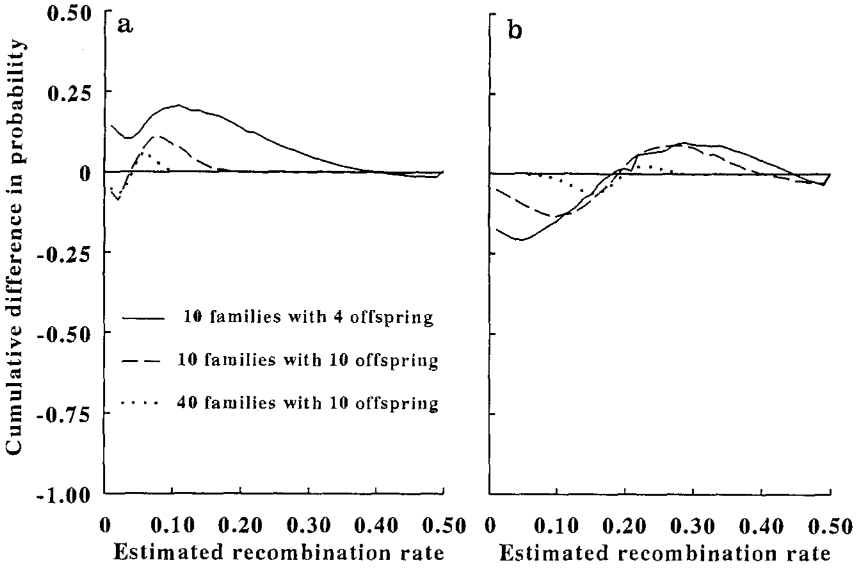

Fig. 2. Cumulative difference between simulated and approximated probability of the *estimated* recombination rate given a *true* recombination rate of (**a**) 0.05 and (**b**) 0.20 for designs with two equiprobable alleles and three sizes

Table 9. Average estimated recombination rate, observed standard error (σ_{obs}), standard error approximated using likelihood function assuming parental phases known (σ_{ap1}), and standard error approximated using correct likelihood function (σ_{ap2}), for designs varying in the number of families, family size and $\hat{\theta}$; one marker has two alleles with equal frequency and one marker locus has six alleles with equal frequency, parental phases are unknown

θ_t	No. obs.[a]	4 offspring per family, 6 observations per family				10[b] offspring per family, 12 observations per family			
		$\hat{\theta}$	σ_{obs}	σ_{ap1}	σ_{ap2}	$\hat{\theta}$	σ_{obs}	σ_{ap1}	σ_{ap2}
0.05	30	0.057	0.083	0.055	0.057	0.052	0.067	0.051	0.051
	60	0.053	0.045	0.039	0.040	0.050	0.039	0.035	0.035
	120	0.051	0.030	0.028	0.028	0.051	0.027	0.025	0.025
	240	0.051	0.021	0.020	0.020	0.049	0.018	0.017	0.018
	480	0.050	0.014	0.014	0.014	0.050	0.013	0.012	0.012
	960	0.050	0.010	0.010	0.010	0.050	0.009	0.009	0.009
0.20	30	0.247	0.166	0.107	0.124	0.219	0.136	0.098	0.102
	60	0.228	0.122	0.076	0.088	0.213	0.094	0.064	0.067
	120	0.210	0.084	0.053	0.062	0.203	0.055	0.048	0.049
	240	0.203	0.049	0.038	0.044	0.202	0.036	0.034	0.035
	480	0.202	0.032	0.027	0.031	0.200	0.025	0.024	0.025
	960	0.201	0.023	0.019	0.022	0.200	0.018	0.017	0.017
0.50	30	0.407	0.140	0.144	*	0.430	0.110	0.131	*
	60	0.414	0.109	0.102	*	0.441	0.086	0.091	*
	120	0.426	0.093	0.072	*	0.447	0.068	0.064	*
	240	0.436	0.077	0.051	*	0.459	0.054	0.045	*
	480	0.449	0.062	0.036	*	0.463	0.046	0.032	*
	960	0.457	0.052	0.025	*	0.470	0.038	0.023	*

[a] No. obs. = number of observations on parents and offspring

[b] For 30 observations, 3 families each with 8 offspring were taken

* σ_{ap2} was undefined (1/0) for θ_t is 0.5

sumption is that, on average, true recombination rate is equal to a given estimated recombination rate. It is not obvious whether or not this assumption is always correct. In Fig. 3 average true recombination rates are plotted against estimated recombination rates. For alternatives with ten families and four or ten offspring per family the average true recombination rate deviated from the given estimated recombination rate. This deviation can be explained by the 0.95 prior-probability that θ_t is 0.5 and the large variance of the estimator when the number of observations on offspring is small.

Most of the deviation between estimated and average true recombination rate disappeared when only replicates were considered for which $\hat{\theta}$ was significantly different

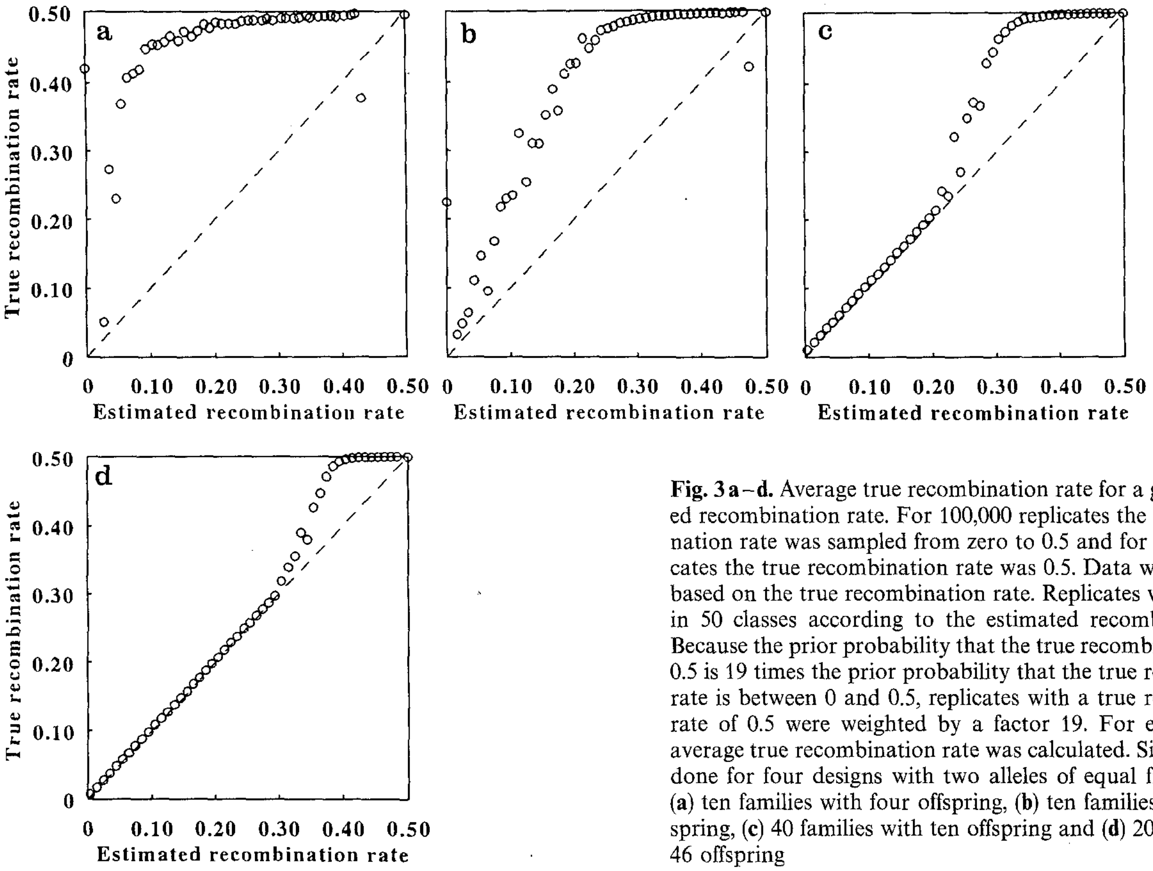

Fig. 3a–d. Average true recombination rate for a given estimated recombination rate. For 100,000 replicates the true recombination rate was sampled from zero to 0.5 and for 100,000 replicates the true recombination rate was 0.5. Data were simulated based on the true recombination rate. Replicates were classified in 50 classes according to the estimated recombination rate. Because the prior probability that the true recombination rate is 0.5 is 19 times the prior probability that the true recombination rate is between 0 and 0.5, replicates with a true recombination rate of 0.5 were weighted by a factor 19. For each class the average true recombination rate was calculated. Simulation was done for four designs with two alleles of equal frequency and (**a**) ten families with four offspring, (**b**) ten families with ten offspring, (**c**) 40 families with ten offspring and (**d**) 20 families with 46 offspring

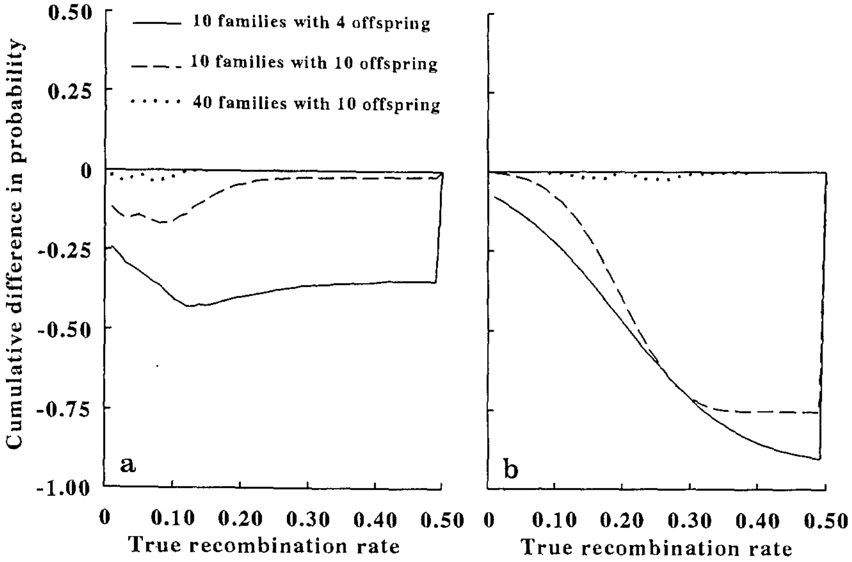

Fig. 4. Cumulative difference between simulated and approximated probability of the *true* recombination rate given an *estimated* recombination rate of (*a*) 0.05 and (**b**) 0.20 for designs with two equiprobable alleles and three sizes

from 0.5, i.e., $Z(\hat{\theta})$ is larger than 3. For the design with ten families and four offspring per family almost no replicates had a $Z(\hat{\theta})$ above 3. This latter observation is consistent with the fact that for large designs the deviation disappeared: recombination rates can only be significant if sufficient observations are available.

The difference between the observed cumulative probability of θ_t given $\hat{\theta}$ and the approximated normal probability is plotted in Fig. 4. In all cases observed cumulative probability was smaller than approximated. With ten families, four offspring per family and a $\hat{\theta}$ of 0.05, for a θ_t of 0.49 the approximated cumulative probability

Table 10. Simulated and approximated probability that true recombination rate is in a specified interval ($0.15 < \theta_t < 0.25$) given that the estimated recombination rate is 0.20

No. families	Family size	No. alleles	P(obs)[a]	P(approx1)[b]	P(approx2)[c]
10	4	2	0.0335	0.1029	0.3507
		6	0.3013	0.6438	0.6359
	10	2	0.1456	0.4600	0.5279
		6	0.8786	0.8543	0.8487
40	10	2	0.8496	0.8529	0.8495
		6	0.9972	0.9963	0.9959
20	46	2	0.9748	0.9721	0.9708
		6	1.0000	0.9999	0.9999

[a] P(obs): observed probability, calculated from simulation (100,000 replicates)

[b] P (approx1): the term $P(\hat{\theta} | 0.15 < \theta_t < 0.25)$ from equation (10) is approximated using the normal distribution

[c] P (approx2): $P\,(0.15 < \theta_t < 0.25 \,|\, \hat{\theta})$ is entirely approximated using the normal distribution

was 0.34 larger than the observed cumulative probability. As expected, no difference was found for a θ_t of 0.5. The observed probability that θ_t is 0.5 for a $\hat{\theta}$ of 0.05 was underestimated by 0.34 by the approximate distribution function. For that design the probability that θ_t is smaller than 0.2 for a $\hat{\theta}$ of 0.05 is overestimated by 0.42 with the approximate distribution. For larger designs this difference was negligible. For a $\hat{\theta}$ of 0.2 probability that θ_t is 0.5 was underestimated by 0.91 using the approximate probability function for designs with ten families of four or ten offspring each (Fig. 4). For the largest design, the approximated cumulative distribution was in good agreement with the observed distribution.

In Table 10 the observed and approximated $P\,(y_1 < \theta_t < y_2 \,|\, \hat{\theta} = x)$ are given. Two approximations were used. In both methods $P(\hat{\theta} = x)$ is calculated as $P\,(x - 0.005 < \hat{\theta} < x + 0.005)$. In the first approximation the probability on an estimate for a given value of θ_t, or $P\,(\hat{\theta} = x \,|\, \theta_t = y)$, is calculated assuming a normal distribution of $\hat{\theta}$ around θ_t. Multiplying $P\,(\hat{\theta} = x \,|\, \theta_t = y)$ by the prior probability of θ_t, integrating over θ_t, and applying equation (10), completes the first approximation. In the second approximation, the true recombination rate is falsely assumed to be normally distributed around $\hat{\theta}$ and an approximation of accuracy is directly obtained from the normal distribution. The second approximation is much more rigorous since the prior probability function of θ_t is ignored. However, Table 10 shows that both approximations worked equally well for the studied alternatives. The first approximation was only better for alternatives for which both approximations were bad. For designs larger than ten families with ten offspring and six equiprobable alleles per locus, approximations were similar and the deviation between the observed and approximated probabilities was small.

Discussion

The expected value of the maximum lod score and the accuracy of an estimated recombination rate were used to describe and study the quality of experimental designs. Ott (1991) argued that the expectation of maximum lod score is not additive over families and has no clear probabilistic interpretation. He concluded that, an expectation of lod score for a given value of θ should be preferred because this expectation is additive over families. The approximated expected maximum lod score used in this study was calculated as the sum of the expected lod score and a constant. Taking the constant into account resulted in a good approximation to the real situation when data from several families were used (Fig. 1).

The approximate methods served two purposes. First, they simplified computations. Second, comprehension of the behavior of estimators was enhanced.

The number of full-sib families, the number of offspring per family and a knowledge of phases were shown to affect $\mathrm{E}[Z(\hat{\theta})]$ (Tables 5 and 8). $\mathrm{E}[Z(\hat{\theta})]$ was larger when the linkage phase was known. The additional number of observations on offspring needed to compensate for lack of knowledge of phases was within a reasonable range (less than six for θ_t smaller than or equal to 0.2 (Table 6). Typing grandparents to determine the parental phase is not an alternative reducing the number of typings to be done for that range of θ_t. However, the additional number of observations on offspring will increase for larger θ_t. For a θ_t larger than 0.20, obtaining information on parental phase might be worthwhile. The aim of most genome mapping projects is to create a map with markers spaced by no more than 20 centimorgans. In such projects the DNA of grandparents is not really needed. Hetzel (1991) pointed out that typing grandparents provides a check for the consistency of segregation. However, typing many offspring also provides a check. The possibility of typing errors emphasized the necessity for typing a large number of offspring per family rather than typing grandparents.

In the present paper, designs with unrelated full-sib families were studied. Elements influencing the quality of designs are most clearly illustrated for this class of designs. Computations are simple. The results for these designs can be used for all other designs with information on parents and offspring when the parental phase is known. In a hierarchical half-sib structure with an equal number of offspring per dam and several dams per sire, fewer sires are used compared to a full-sib structure with the same number of dams and offspring. When parental phases are unknown, fewer sires means that less information will be used to infer parental phases from the data. As a consequence $\mathrm{E}[Z(\hat{\theta})]$ will be larger for the hierarchical half-sib structure.

Results in Table 8 showed that $E[Z(\hat{\theta})]$ can be maximized by minimizing the number of families. A minimum number of families is not necessarily optimal, however. With a minimum number of families the variation in realized maximum lod score is maximal and, as a consequence, the probability of having no information is maximal. The risk due to a large variation in the outcome of an experiment can be summarized by the probability of no information. Assume the probability of no information is to be less than 0.10. The necessary number of families can be calculated as $-1/\log_{10}[P(\text{type I})]$ where $P(\text{type I})$ is the probability that a family is of type I. For a design where marker loci have two equiprobable alleles, $P(\text{type I})$ is 0.5625 and the probability of no information is 0.1 when the number of families is four. Given this number of families, the number of offspring per family resulting in $E[Z(\hat{\theta})] = 3$ can be calculated. When marker loci have two equiprobable alleles, four families each with 33 offspring are needed for an $E[Z(\hat{\theta})]$ of 3. When marker loci have four equiprobable alleles the probability of no information is 0.04 with two families and in that case 21 offspring per family are needed for a $E[Z(\hat{\theta})]$ of 3. The number of offspring per full-sib family in these examples is larger than that available in most livestock species. The restriction on probability of no information will not change the optimal design when the number of offspring per family is less than 20. There is, however, still a risk that the realized maximum lod score in an experiment is lower than $E[Z(\hat{\theta})]$. A more general approach is to look at the power of a design.

The relation between marker polymorphism and distribution over family types demonstrated the advantage of highly polymorphic markers (Table 4). The research of Georges et al. (1990) in cattle showed an average heterozygosity of 51% for VNTR markers and 65% for microsatellites. These heterozygosities correspond to about two or three alleles of equal frequency (see Table 4), or more alleles of varying frequencies. Consequently, on average, a considerable proportion of the families will provide no, or less than maximal, information on linkage.

Boehnke (1986) described a simulation approach by which the average maximum lod scores and power can be obtained for any design. For plants, elements of the design of experiments are described in standard text books (e.g. Mather 1951; Bailey 1961; Green 1981). Restriction is usually made to designs with double backcrosses or intercrosses, family types III and V respectively, and known phases. The present study described all possible family types in a segregating population and paid attention to families larger than those used for human linkage studies. The derived algorithm considers marker polymorphism, all family types, and can be used for varying full-sib family sizes and the number of families. The method of Boehnke (1986) is general and further calculates power but requires simulation of many replicates for each design to be evaluated.

Accuracy was calculated from the distribution function of θ_t for a given value of $\hat{\theta}$. Elements of the function were studied and compared with approximations. Bias in the estimate of recombination rate, given a θ_t for designs with few observations, could be explained by the observed distribution of the estimates. Bias of estimated recombination rates was studied in more detail by Bolling and Murphy (1979). For designs larger than or equal to ten families with ten offspring, use can be made of the normal distribution with an approximated standard error to calculate the probability for an estimate given a true recombination value.

The use of the prior-probability density of θ_t in linkage studies has been advocated by Smith (1959), Smith and Sturt (1976), Silver and Buckler (1986) and Neumann (1990, 1991). This approach considers the 0.95 probability that loci are unlinked. The effect of prior density of true recombination rate is shown in Fig. 3. The 0.95 probability of no linkage resulted in a large deviation between true and estimated recombination rate. The influence of the large probability of unlinked loci on average estimated recombination rate could be reduced by considering only replicates which had an estimate for θ significantly different from 0.5.

For large designs, inferences about true recombination rate can be made using the normal distribution and the approximated standard error. For small designs, a restriction needs to be made to significant recombination rates. The results of this study emphasize the necessity to use significant estimates because non-significant estimates are not only inaccurate but, on average, are very different from true recombination rates.

Acknowledgements. The authors acknowledge financial support from Euribrid B.V. They also thank two anonymous reviewers for their useful comments.

References

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

Bitgood JJ, Somes RG (1990) Linkage relationships and gene mapping. In: Crawford RD (ed) Poultry breeding and genetics. Elsevier, Amsterdam, pp 469–495

Boehnke M (1986) Estimating the power of a proposed linkage study: a practical computer simulation approach. Am J Hum Genet 39:513–527

Bolling DR, Murphy EA (1979) Finite sample properties of maximum likelihood estimates of the recombination fraction in double backcross matings in man. Am J Med Genet 3:81–95

Brascamp EW, Haley CS, Fries R, Grosclause F, Hanset R, Womack J (1991) Genome mapping in farm animals, a survey of current research. Paper presented at the 42nd annual meeting of the EAAP, Berlin

Fries R, Beckmann JS, Georges M, Soller M, Womack J (1989) The bovine gene map. Anim Genet 20:3–29

Georges M, Mishra A, Sargeant L, Steele M, Zhao X (1990) Progress towards a primary DNA marker map in cattle. Proc 4th World Congr Genet Appl Livestock Prod, Edinburgh, 13:107–112

Green EL (1981) Genetics and probability in animal breeding experiments. Oxford University Press, New York

Haldane JBS (1919) The combination of linkage values, and the calculation of distance between the loci of linked factors. J Genet 8:299–309

Haley CS, Archibald A, Andersson L, Bosma AA, Davies W, Fredholm M, Geldermann H, Groenen M, Gustavsson I, Ollivier L, Tucker EM, Van de Weghe A (1990) The pig gene mapping project – PiGMaP. Proc 4th World Congr Genet Appl Livestock Prod, Edinburgh, 13:67–70

Hetzel DJS (1991) Reference families for genome mapping in domestic livestock. In: Schook LB, Lewin HA, McLaren DG (eds) Gene-mapping techniques and applications. Marcel Dekker, New York, pp 51–64

Kendall M, Stuart A (1978) The advanced theory of statistics, vol 2. Griffin, London

Kennedy BW, Verrinder Gibbins AM, Gibson JP, Smith C (1990) Coalescence of molecular and quantitative genetics for livestock improvement. J Dairy Sci 73:2619–2627

Mather K (1951) The measurement of linkage in heredity. Methuen, London

Morton NE (1955) Sequential test for the detection of linkage. Am J Hum Genet 7:277–318

Neuman PE (1990) Two-locus linkage analysis using recombinant inbred strains and Bayes' theorem. Genetics 126:277–284

Neumann PE (1991) Three-locus linkage analysis using recombinant inbred strains and Bayes' theorem. Genetics 128: 631–638

Ott J (1991) Analysis of human genetic linkage. The Johns Hopkins University Press, London

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

Silver J, Buckler CE (1986) Statistical considerations for linkage analysis using recombinant inbred strains and backcrosses. Proc Natl Acad Sci USA 83:1423–1427

Smith CAB (1959) Some comments on the statistical methods used in linkage investigations. Am J Hum Genet 11:289–304

Smith CAB, Sturt E (1976) The peak for the likelihood curve in linkage testing. Ann Hum Genet 39:423–426

Soller M, Beckmann JS (1983) Genetic polymorphism in varietal identification and genetic improvement. Theor Appl Genet 67:25–33