Theor Appl Genet (2003) 106:840–845
DOI 10.1007/s00122-002-1099-6

M. Hühn · H. P. Piepho

Determining the sample size for co-dominant molecular marker-assisted linkage detection for a monogenic qualitative trait by controlling the type-I and type-II errors in a segregating F2 population

Received: 11 February 2002 / Accepted: 6 August 2002 / Published online: 19 December 2002

Abstract Tests for linkage are usually performed using the lod score method. A critical question in linkage analyses is the choice of sample size. The appropriate sample size depends on the desired type-I error and power of the test. This paper investigates the exact type-I error and power of the lod score method in a segregating F_2 population with co-dominant markers and a qualitative monogenic dominant-recessive trait. For illustration, a disease-resistance trait is considered, where the susceptible allele is recessive. A procedure is suggested for finding the appropriate sample size. It is shown that recessive plants have about twice the information content of dominant plants, so the former should be preferred for linkage detection. In some cases the exact α-values for a given nominal α may be rather small due to the discrete nature of the sampling distribution in small samples. We show that a gain in power is possible by using exact methods.

Keywords Power · Type-I error · Maximum likelihood · Fisher information · Co-dominant markers · LOD score · Linkage analysis · Sample size

Introduction

Because of their almost unlimited number and due to their independence from environmental factors as well as dominance and epistatic effects, co-dominant molecular markers are highly superior to biochemical markers and morphological markers. Linkage maps, therefore, have been constructed for many economically important organisms (Sunil 1999). In the field of practical plant breeding, for example, the selection for a dominant disease resistance trait is a common breeding objective (e.g. nematode resistance in sugar beet, Jung et al. 1992). Linkage maps of crop species are often constructed with segregating populations, i.e., F_2 populations or backcrosses (Sunil 1999).

In this paper, we present our results on necessary sample sizes for molecular marker-assisted linkage detection for a dichotomous (dominant-recessive) trait. Analysis is based on a segregating F_2 population. The sample size determination is carried out by controlling type-I and type-II errors [type-I error = α = probability that the test rejects H_0 (no linkage) although H_0 is true; type-II error = β = probability that the test fails to reject H_0 although H_0 is false]. It is usually convenient to work not with the type-II error (β) but rather with its complement, the power ($1 - \beta$). The power ($1 - \beta$) is the probability that a test rejects H_0 when it is false; i.e., the probability of detecting linkage when it exists.

In most linkage analyses an explicit consideration of both types of error is ignored and the actual numerical magnitudes of α and β are completely unknown. In many tests of genetic linkage, only type-I errors are considered and type-II errors are commonly ignored. More often than not, however, it is the power of the test for linkage that is of primary importance.

By the traditional 'lod score' method of linkage analysis, the test for linkage of autosomal loci is declared significant when the maximum of the lod score (Z) exceeds the bound $Z_0 = 3.0$ (Ott 1991). The lod score is essentially a likelihood ratio statistic for the test of no linkage. The rationale for the bound $Z_0 = 3$ lies in the fact that for the likelihood ratio test $\alpha \leq 10^{-z_0}$ (Ott 1991), so for $Z_0 = 3$ we have $\alpha \leq 0.001$. Unfortunately, this upper bound for the type-I error may be rather conservative

Communicated by H.C. Becker

Electronic Supplementary Material Supplementary material is available for this article if you access the article at http://dx.doi.org/10.1007/s00122-002-1099-6. A link in the frame on the left on that page takes you directly to the supplementary material.

M. Hühn
Institute for Crop Science and Plant Breeding, University of Kiel, Olshausenstrasse 40, 24118 Kiel, Germany,

H. P. Piepho (✉)
Institute for Crop Production and Grassland Research, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany,
e-mail: piepho@uni-hohenheim.de

(Ott 1991), so that the use of the true $\alpha = 0.001$ threshold would usually increase power.

This paper employs exact methods that allow us to assess type-I error and the power of tests for linkage and to choose an appropriate sample size. All of the investigations reported in this paper are restricted to the analysis of traditional two-point linkage (Ott 1991).

Problem and theory

The investigations are based on a diploid segregating F_2 population co-segregating for a molecular marker and a gene coding for a qualitative trait. For illustrative purposes, the case of a disease resistance gene will be considered here, where the susceptible allele is recessive, while the resistant allele is dominant. In this case, we suggest that linkage analysis be based on susceptible plants. The situation in which the susceptible allele is dominant is completely analogous and is identical to the case presented here, with the only difference that linkage analysis is based on resistant plants. The results given here are generally applicable for a qualitative trait with one recessive and one dominant allele. The two alleles at the resistance gene locus are denoted by A (= resistant) and a (= susceptible) with A dominant over a. The marker alleles with co-dominant expression are B_1 and B_2 with a recombination value R between the marker and the disease resistance gene locus.

Selfing or intercrossing the F_1 generation AaB_1B_2 of an initial cross of homozygous parents creates a segregating F_2 population. In this paper, linkage analysis is based on the sub-population of susceptible (= recessive) individuals of this F_2. Results for the sub-population of resistant individuals and the total population are available from the second author upon request. The double heterozygote AaB_1B_2 produces the gametes AB_1, aB_1, AB_2 and aB_2 with frequencies $^1/_2(1 - R)$, $^1/_2R$, $^1/_2R$ and $^1/_2(1 - R)$, respectively. The recombination value is assumed to be equal in both sexes. The composition of the segregating F_2 is given in Table 1.

The approach of this paper is restricted to the analysis of the simplest situation of two-point linkage by the traditional approach of maximum lod score. This is defined as the logarithm of base 10 of the ratio of the likelihoods when the loci are at their maximum-likelihood recombination fraction and when the loci are taken to be unlinked.

We denote:

- N = number of tested individuals,
- k = number of phenotypically distinct classes (considering both the resistance and the marker loci),
- f_i = expected relative frequency of class
- z_i = observed absolute frequency of class
- $L(R)$ = likelihood function dependent on the recombination fraction
- $Z(R)$ = lod score for recombination value

Table 1 Genotypes of a segregating F_2 population from a cross $AAB_1B_1 \times aaB_2B_2$ with associated gametic and genotypic frequencies

Female gametes with frequencies \ Male gametes with frequencies	AB_1 $\frac{1}{2}(1-R)$	AB_2 $\frac{1}{2}R$	aB_1 $\frac{1}{2}R$	aB_2 $\frac{1}{2}(1-R)$
AB_1 $\frac{1}{2}(1-R)$	AAB_1B_1 $\frac{1}{4}(1-R)^2$	AAB_1B_2 $\frac{1}{4}R(1-R)$	AaB_1B_1 $\frac{1}{4}R(1-R)$	AaB_1B_2 $\frac{1}{4}(1-R)^2$
AB_2 $\frac{1}{2}R$	AAB_1B_2 $\frac{1}{4}R(1-R)$	AAB_2B_2 $\frac{1}{4}R^2$	AaB_1B_2 $\frac{1}{4}R^2$	AaB_2B_2 $\frac{1}{4}R(1-R)$
aB_1 $\frac{1}{2}R$	AaB_1B_1 $\frac{1}{4}R(1-R)$	AaB_1B_2 $\frac{1}{4}R^2$	aaB_1B_1 $\frac{1}{4}R^2$	aaB_1B_2 $\frac{1}{4}R(1-R)$
aB_2 $\frac{1}{2}(1-R)$	AaB_1B_2 $\frac{1}{4}(1-R)^2$	AaB_2B_2 $\frac{1}{4}R(1-R)$	aaB_1B_2 $\frac{1}{4}R(1-R)$	aaB_2B_2 $\frac{1}{4}(1-R)^2$

The maximum likelihood estimate $\hat{R}$ of the recombination fraction R is that value for which the lod score (or equivalently, the likelihood or log-likelihood) is maximized. A conventional rule is to conclude that autosomal loci are linked whenever the maximum lod score exceeds 3 (Ott 1991). What are the associated true type-I errors of this approach? The upper bound for the type-I error is 0.001, but the true type-I error may be much lower. To compute the exact type-I error and the power, exact distributions of the lod score need to be considered. Since the lod score depends on the number N of tested individuals, the exact distributions of $\hat{R}$ and of $Z(\hat{R})$ under H_0 must be calculated for different values of N.

In many cases, a "cook-book"-type application of the conventional critical lod score values (2, 3 or other) actually involves true type-I and type-II errors of usually unknown magnitudes. Determination of necessary sample sizes in linkage analysis based on the exact distributions and controlling prespecified type-I and type-II errors, therefore, should be preferred. Based on the exact distributions of $\hat{R}$ and $Z(\hat{R})$ under H_0 and H_1 for the subpopulation of susceptible individuals, the required sample size can be easily calculated by controlling type-I and type-II errors. We consider the null hypothesis H_0: $R = 0.5$ and the alternative H_1: linkage with true recombination fraction R.

We suggest basing linkage analyses only on the subpopulation of susceptible (recessive) individuals of the F_2, since these have the highest information content. This is so, because there are only three susceptible genotypes at the marker level, which can all be distinguished phenotypically, whereas not all resistant genotypes can be separated based on their phenotype. A more detailed discussion is given in the next section.

The expected relative frequencies f_i, $(i = 1, 2, \ldots, k)$ of the three phenotypically distinct classes $(k = 3)$ of susceptible (recessive) plants are (Table 1):

$$f_1 = R^2 \quad (\text{for } aaB_1B_1),$$
$$f_2 = 2R(1-R) \quad (\text{for } aaB_1B_2), \tag{1}$$
$$f_3 = (1-R)^2 \quad (\text{for } aaB_2B_2).$$

The observed absolute frequencies are denoted as z_1, z_2 and z_3 respectively, with $z_1 + z_2 + z_3 = N$ = number of tested susceptible individuals. The likelihood function $L(R)$ is

$$L(R) = 2^{z_2} R^{2z_1+z_2}(1-R)^{z_2+2z_3}. \tag{2}$$

With $L(0.5) = 2^{z_2-2N}$ $L(0.5) = 2^{z2-2N}$ one obtains for the lod score

$$Z(R) = \log_{10}[R^{2z_1+z_2}(1-R)^{z_2+2z_3}2^{2N}] \tag{3}$$

The maximum likelihood estimate $\hat{R}$ is found by setting to zero the first derivative of the lod score (or equivalently of the log-likelihood) and verifying that this maximizes the lod score. We find the following estimate:

$$\hat{R} = \frac{2z_1 + z_2}{2N}.$$

For known coupling phase (which is assumed throughout this paper), R is restricted to lie between 0 and 0.5. Thus, estimates of R larger than 0.5 may be set equal to 0.5:

$$\hat{R} = \frac{2z_1 + z_2}{2N} \quad \text{if } \frac{2z_1 + z_2}{2N} < 0.5,$$
$$\hat{R} = 0.5 \quad \text{if } \frac{2z_1 + z_2}{2N} \geq 0.5. \tag{4}$$

For the maximum lod score one obtains:

$$Z(\hat{R}) = \log_{10}(2^{2N}) \quad \text{if} \frac{2z_1 + z_2}{2N} = 0,$$
$$Z(\hat{R}) = \log_{10}\left[(2z_1 + z_2)^{2z_1+z_2}(z_2 + 2z_3)^{z_2+2z_3}/N^{2N}\right]$$
$$\text{if} 0 < \frac{2z_1 + z_2}{2N} < 0.5,$$
$$Z(\hat{R}) = 0 \quad \text{if} \frac{2z_1 + z_2}{2N} \geq 0.5. \tag{5}$$

The type-I and type-II errors are derived from the exact distributions of $\hat{R}$ and $Z(\hat{R})$ under the null hypothesis H_0: no linkage (i.e. $R = 0.5$) and under H_1: linkage with a specified recombination value R ($0 \leq R < 0.5$). The exact distributions of $\hat{R}$ and $Z(\hat{R})$ under H_0 and under H_1 can be easily found by noting that (z_1, z_2, z_3) is multinomially distributed with parameters N, $f_1 = R^2$, $f_2 = 2R(1–R)$, and $f_3 = (1– R)^2$. Since $Z(\hat{R})$ is a monotonically decreasing function of $\hat{R}$ on the interval [0; 0.5], both statistics are equivalent and it suffices to consider the distribution of $\hat{R}$.

In order to obtain the exact distributions of $\hat{R}$ we use the following approach:

Step 1: For all possible combinations of (z_1, z_2, z_3) compute $\hat{R}$ by Eq. 4.

Step 2: Sort $\hat{R}$-values in ascending order and from the associated (z_1, z_2, z_3)-values compute the corresponding multinomial probability.

Step 3: Sum up the probabilities for equal $\hat{R}$-values.

Step 4: Calculate the cumulative probabilities.

The required sample size for a type-I error α, a power of 1 – β and linkage with a recombination value $R < 0.5$ is obtained by the following computational procedure:

1. Set N= 2.
2. Compute cumulative probabilities for R=0.5 and R<0.5.
3. From the distribution for R= 0.5, determine the largest value of $\hat{R}$ for which the cumulative probability is smaller or equal to the specified α. This value of $\hat{R}$ will be denoted as R_{crit}; the corresponding probability will be denoted as α_{exact}. In the test, values $\hat{R} \leq R_{crit}$ are judged to be significant. Determine the cumulative probability of the distribution with an R< 0.5 at a critical level R_{crit}. This cumulative probability is the exact power of the test. It is denoted as $(1 - \beta_{exact})$. If the power $(1 - \beta_{exact})$ is smaller than the predetermined power $(1 – \beta)$, then augment $N$by one and go to step (2). The final value of $N$is the required sample size for a type-I error α, a power of $(1 – \beta)$ and a linkage with recombination value R. Note that usually $\alpha > \alpha_{exact}$and $(1 – \beta) < (1 – \beta_{exact})$.

Results and discussion

Numerical results are presented in Table 2 for a Type I error of $\alpha = 0.001$. The computations were programmed using the SAS system. The program is easily adapted for other values of the power, the type-I error α and the sample size N. Tables for other values of α as well as the SAS source code are available as Electronic Suplementary Material at http://dx.doi.org/10.1007/s00122-002-1099-6. Suppose the researcher wants to detect linkages $R = 0.05$ or smaller with a power of 0.95 or larger at a type-I error rate of $\alpha = 0.001$. From Table 2 we find that the necessary sample size is $N = 11$. At this sample size, the power is $(1 – \beta_{exact}) = 0.98$ and the type-I error is $\alpha_{exact} = 0.00042772$.

For small N, the distribution of $\hat{R}$ is pronouncedly discrete, so α_{exact} varies notably with N. The variation in α_{exact} may be quite substantial even for relatively large N. In a given experiment, it may be useful to consider the two values of α_{exact} immediately below and above the value of α_{exact} closest to the intended α. In this paper, the simplest case was presented, where only recessive (susceptible) genotypes are used for linkage analysis. This was done because these genotypes have the highest information content. In practice, the major share of the total cost of a linkage analysis is spent on marker identification, while raising the plants and subjecting them to resistance testing is not usually a limiting factor. Therefore, it is often affordable to raise resistant plants together with susceptible ones and then select only susceptible ones for linkage analysis.

Some explanation as to why susceptible plants have a higher information content than resistant plants is in order.

Table 2 Exact critical values of $\hat{R}$ (R_{crit}), exact type-I errors (α_{exact}) and exact power for different values of R at an upper bound α = 0.001. Values $\hat{R} \leq R_{crit}$ are judged to be significant

Sample size	R_{crit}	Power						α_{exact}
N		R = 0.01	R = 0.02	R = 0.05	R = 0.10	R = 0.20	R = 0.30	
5	0.00000	0.90	0.82	0.60	0.35	0.11	0.03	0.00097656
6	0.00000	0.89	0.78	0.54	0.28	0.07	0.01	0.00024414
7	0.07143	0.99	0.97	0.85	0.58	0.20	0.05	0.00091553
8	0.06250	0.99	0.96	0.81	0.51	0.14	0.03	0.00025940
9	0.11111	1.00	0.99	0.94	0.73	0.27	0.06	0.00065613
10	0.10000	1.00	0.99	0.92	0.68	0.21	0.04	0.00020123
11	0.13636	1.00	1.00	0.98	0.83	0.33	0.07	0.00042772
12	0.16667	1.00	1.00	0.99	0.91	0.46	0.11	0.00077194
13	0.15385	1.00	1.00	0.99	0.89	0.38	0.07	0.00026676
14	0.17857	1.00	1.00	1.00	0.94	0.50	0.11	0.00045612
15	0.20000	1.00	1.00	1.00	0.97	0.61	0.16	0.00071545
16	0.18750	1.00	1.00	1.00	0.96	0.54	0.11	0.00026753
17	0.20588	1.00	1.00	1.00	0.98	0.63	0.16	0.00041070
18	0.22222	1.00	1.00	1.00	0.99	0.72	0.20	0.00059662
19	0.23684	1.00	1.00	1.00	1.00	0.78	0.26	0.00082903
20	0.22500	1.00	1.00	1.00	0.99	0.73	0.20	0.00033977
21	0.23810	1.00	1.00	1.00	1.00	0.80	0.24	0.00047034
22	0.25000	1.00	1.00	1.00	1.00	0.85	0.29	0.00063002
23	0.26087	1.00	1.00	1.00	1.00	0.89	0.34	0.00082075
24	0.25000	1.00	1.00	1.00	1.00	0.85	0.28	0.00035863
25	0.26000	1.00	1.00	1.00	1.00	0.89	0.33	0.00046811
26	0.26923	1.00	1.00	1.00	1.00	0.92	0.38	0.00059756
27	0.27778	1.00	1.00	1.00	1.00	0.94	0.43	0.00074813
28	0.28571	1.00	1.00	1.00	1.00	0.96	0.47	0.00092078
29	0.27586	1.00	1.00	1.00	1.00	0.94	0.41	0.00043089
30	0.28333	1.00	1.00	1.00	1.00	0.96	0.45	0.00053288
31	0.29032	1.00	1.00	1.00	1.00	0.97	0.50	0.00064950
32	0.29688	1.00	1.00	1.00	1.00	0.98	0.54	0.00078139
33	0.30303	1.00	1.00	1.00	1.00	0.98	0.58	0.00092912
34	0.29412	1.00	1.00	1.00	1.00	0.98	0.52	0.00045720
35	0.30000	1.00	1.00	1.00	1.00	0.98	0.56	0.00054662
40	0.31250	1.00	1.00	1.00	1.00	0.99	0.65	0.00052637
45	0.32222	1.00	1.00	1.00	1.00	1.00	0.72	0.00048638
50	0.34000	1.00	1.00	1.00	1.00	1.00	0.84	0.00089497
60	0.35000	1.00	1.00	1.00	1.00	1.00	0.90	0.00064967
70	0.36429	1.00	1.00	1.00	1.00	1.00	0.96	0.00083376
80	0.37500	1.00	1.00	1.00	1.00	1.00	0.98	0.00097686
90	0.37778	1.00	1.00	1.00	1.00	1.00	0.99	0.00064302
100	0.38500	1.00	1.00	1.00	1.00	1.00	1.00	0.00070085
120	0.39583	1.00	1.00	1.00	1.00	1.00	1.00	0.00075443
140	0.40357	1.00	1.00	1.00	1.00	1.00	1.00	0.00074665
160	0.40938	1.00	1.00	1.00	1.00	1.00	1.00	0.00070139

Among susceptible plants, there are only three genotypes, and these can all be distinguished phenotypically, whereas there are six different resistant genotypes but only three distinct phenotypes. From this fact, a higher information content of susceptible plants is expected. A more rigid explanation can be based on Fisher's information, i.e. on the second derivative of the log-likelihood with respect to the parameter R. Fisher's information $i(R)$ is given by

$$i(R) = -\left[\frac{\partial^2 \log L}{\partial R^2}\right].$$

In large samples, the variance of the maximum likelihood estimate of R is (Weir 1996)

$$Var(\hat{R}) = \frac{1}{E\,[i(R)]}$$

where $E[\cdot]$ denotes the expected value. The expected information $E[i(R)]$ may be used to compare three different settings: (1) only susceptible plants used; (2) only resistant plants used; (3) all plants (susceptible and resistant) used. For this comparison, it is useful to consider the expected information per experimental unit (plant): $E[i(R)]/N$. The explicit expressions for the expected information per plant are (Hühn 1995):

$2/(R-R^2)$ for susceptible plants,

$\frac{1}{3}[4/(1-R^2)+4/(2R-R^2)-6/(1+R^2-R)]$

for resistant plants,

$[1/(1-R^2)+1/(2R-R^2)+1/(2R-2R^2)-3/(2R^2-2R+2)]$ for the complete F_2.

Figure 1 shows the dependence of the expected information per plant on the recombination rate R and the plant material. It can be seen that the susceptible plants have the highest information content.

Sample size calculations for the cases of resistant plants and of all plants can be performed in the same way as in the case of susceptible plants. The only difference is that maximum likelihood estimates of the recombination rate must be obtained numerically (Weber and Wricke 1994), e.g., by a Newton-Raphson algorithm (Weir 1996) or by an EM algorithm (Ott 1977). A computer program based on a Newton-Raphson algorithm (using the SAS system) is available as Electronic Suplementary Material at http://dx.doi.org/10.1007/s00122-002-1099-6.

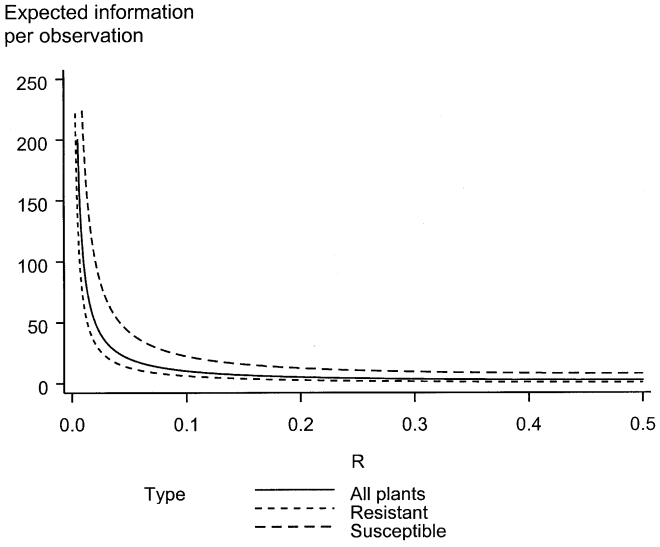

Fig. 1 Information content per plant (expected Fisher information E[$i(R)$], divided by sample size N) depending on genetic composition of plant material (all plants, resistant plants only, susceptible plants only) and true recombination rate

Determining α

Morton (1955) states that one should be "especially anxious to avoid the assertion that two genes are linked when in fact they are not, since a misleading linkage map is worse than no linkage map at all." This suggests that a very small α, like 0.0001, is appropriate. A more rigid justification for small α is this: cases of apparent linkage will be made up in part of true linkages, in part of type-I errors. What the experimenter usually needs to control is the posterior probability θ that an apparent linkage observed among numerous markers and a gene of interest is a type-I error rather than a true linkage. This posterior type-I error probability θ clearly is not the same as α, but it is related to α. This sub-section describes a strategy for chosing α so that the posterior probability θ is controlled at the pre-specified level.

To compute the posterior probability θ it is necessary to have an idea of the probability of linkage among two loci (ϕ). This probability is roughly equal to the probability that two randomly chosen loci are on the same chromosome, i.e. $\phi \sim h^{-1}$, where h is the haploid number of chromosomes (Morton 1955). A more accurate assessment of ϕ would need to account for the length of all chromosomes, because two loci on the same chromosomes are unlinked when $R = 0.5$ (see comments below). Here, we will just assume $\phi \sim h^{-1}$ for simplicity. Another quantity, that needs to be determined is the average power of the used test to detect linkage when, in fact, the two loci are linked. This average will be denoted as $\bar{P}$. With this information, θ is given by (Morton 1955; Ott 1991)

$$\theta = \frac{\alpha(1-\phi)}{\alpha(1-\phi)+\phi\bar{P}}. \tag{6}$$

Solving Eq. (6) for α leads to

$$\alpha = \frac{\theta\phi\bar{P}}{(1-\phi)(1-\theta)} \tag{7}$$

To illustrate Eq. (6) assume that the probability that two randomly chosen loci are linked is $\phi = 0.05$. Also assume that the power of the test for linkage is always exactly 100%, i.e., one will detect every linkage whenever it is present. If we choose a type-I error of $\alpha = 0.05$, 48.7% of our significant tests will be false positives, i.e. $\theta = 0.487$ (Ott 1991). Equation 7 may be used to determine a more appropriate α. If, for example the average power is $\bar{P} = 0.95$, the probability of linkage is $\phi = 0.07$ and the experimenter wants to control the posterior probability of a type-I error at $\theta = 0.05$, α should be chosen smaller than 0.004.

For linked loci, R may be regarded as uniformly distributed between 0 and 1/2 (Morton 1955), hence

$$\bar{P} = 2\int_0^{1/2} P(R)dR$$

where $P(R)$ is the power for a given value of R. From the exact distribution of $\hat{R}$ for given R, and the a priori distribution of R, it is easy to derive $P(R)$. Integration may be approximately done by averaging $P(R)$ for a dense grid of R-values, e.g. 0.001, 0.002, 0.003, …, 0.5.

It is noted that $P(R)$ depends on α and N. One can either fix α and choose the smallest N for which the desired power $\bar{P}$ is attained, or one fixes N and finds the largest α that yields the desired power $\bar{P}$. When fixing α, the upper bound

$$\alpha < \frac{\theta\phi}{(1-\phi)(1-\theta)} \tag{8}$$

should be observed. Some numerical values for these upper bounds are given in Table 3. For probabilities θ and ϕ smaller than 10%, the true type-I error rate α must be smaller than 1%; i.e. the true values of α are particularly small! Since $P(R)$ depends on α and N, Eq. (7) usually

Table 3 Numerical values for the upper bound on α in Eq. (8)

θ	ϕ				
	0.02	0.04	0.06	0.08	0.10
0.02	0.0004	0.0009	0.0013	0.0018	0.0023
0.04	0.0009	0.0017	0.0027	0.0036	0.0046
0.06	0.0013	0.0027	0.0041	0.0056	0.0071
0.08	0.0018	0.0036	0.0056	0.0076	0.0097
0.10	0.0023	0.0046	0.0071	0.0097	0.0123

Table 4 Exact Type-I errors for different critical values of the lod score and different values of *N* (sample size)

N	lod = 2	lod = 3	lod = 4
4	0.0039063	0.00000000	0.000000000
6	0.0031738	0.00024414	0.000000000
8	0.0020905	0.00025940	0.000015259
10	0.0012884	0.00020123	0.000020027
12	0.0007719	0.00013858	0.000017941
14	0.0018596	0.00009000	0.000013720
16	0.0010512	0.00005654	0.000009651
18	0.0019666	0.00015628	0.000006457
20	0.0011107	0.00009108	0.000004182
25	0.0013011	0.00015293	0.000011931
30	0.0013352	0.00006726	0.000006073
35	0.0012738	0.00008302	0.000009600
40	0.0011623	0.00009156	0.000012646
45	0.0010301	0.00009388	0.000005451
50	0.0008950	0.00009157	0.000006290
Upper bound	0.0100000	0.00100000	0.000100000

has to be solved iteratively. Moreover, N is the sample size that needs to be determined for given α, β and R. So the following iterative procedure seems reasonable:

1. Determine θ, ϕ, and β and choose reasonable starting values for α and N.
2. Compute $\bar{P}$ for β and the current values of α and N.
3. Compute α from Eq. (7).
4. Determine exact sample size N for given θ, ϕ, and β and current α.
5. Iterate steps (2) through (4) until $N$and α converge.

One problem with this computational procedure is to find a good starting value for N. One option is to use the normal approximation for sample size in the one-sided one-sample normal means problem (Steel and Torrie 1980). Adapted to our problem, this yields $N = \mathrm{var}(\hat{R})(z_{1-\alpha} + z_{1-\beta})^2/(0.50 - R)^2$, where $z_{1-\alpha}$ is the (1–α)-percentile of the standard normal distribution.

Our approximate approach for computing θ is based on the simplifying assumption that $\phi \sim h^{-1}$. For an exact assessment, the lengths of the different linkage groups need to be taken into account. A simple approach is to simulate the a priori distribution of R by repeatedly sampling independent pairs of loci from the whole genome and determining their recombination fraction R. This simulated a priori distribution of R may be used to derive both ϕ and $\bar{P}$. Up to simulation errors, the resulting θ will be exact. Computations are straightforward, but computationally more demanding than the approximate method.

Finally, some numerical results on true type-I errors for critical lod scores 2, 3 and 4 are given in Table 4. The exact values are considerably below the upper bound, which shows that the traditional lod score method is rather conservative in the case at hand, and a gain in power is possible by using exact methods.

This paper has presented a method to compute exact type-I and type-II errors for the lod score method to detect linkage between a qualitative trait and a co-dominant marker. We have suggested choosing the type-I error rate, α, so that the a posterori probability of linkage, θ, given a significant test, is controlled at a pre-specified value. The latter probability is quite relevant, since one would like to be confident that a significant linkage is not a false positive. Control of the probability of false positives among detections is available if some a priori assumptions can be made about the probability of linkage.

References

Hühn M (1995) Determining the linkage of disease-resistance genes to molecular markers: the LOD-SCORE method revisited with regard to necessary sample sizes. Theor Appl Genet 90:841–846

Jung C, Koch R, Fischer F, Brandes A, Wricke G, Herrmann RG (1992) DNA markers closely linked to nematode resistance genes in sugar beet (Beta vulgaris L.) using chromosome additions and translocations originating from wild beets of the Procumbentes species. Mol Gen Genet 232:271–278

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

Ott J (1977) Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. Ann Hum Genet 40:442–454

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

Steel RGD, Torrie JH (1980) Principles and procedures of statistics: a biometrical approach. McGraw-Hill, New York

Sunil KL (1999) DNA markers in plant improvement: an overview. Biotechnol Adv 17:143–182

Weber WE, Wricke G (1994) Genetic markers in plant breeding. Advances in plant breeding, vol 16 (supplements to the journal Plant Breeding). Paul Parey, Berlin Hamburg

Weir BS (1996) Genetic data analysis II. Sinauer, Sunderland