Theor Appl Genet (1994) 88:293–297

S. Gerber · F. Rodolphe

Estimation and test for linkage between markers: a comparison of lod score and χ^2 test in a linkage study of maritime pine (*Pinus pinaster* Ait.)

Received: 15 August 1993 / Accepted: 9 September 1993

Abstract The first step in the construction of a linkage map involves the estimation and test for linkage between all possible pairs of markers. The lod score method is used in many linkage studies for the latter purpose. In contrast with classical statistical tests, this method does not rely on the choice of a first-type error level. We thus provide a comparison between the lod score and a χ^2 test on linkage data from a gymnosperm, the maritime pine. The lod score appears to be a very conservative test with the usual thresholds. Its severity depends on the type of data used.

Key words Linkage test · Lod score · Pine

Introduction

The second law of Mendel asserts that when two or more pairs of characters are involved in a cross, their relevant segregation occurs independently. However, exceptions to this law had been discovered as early as 1902. The basis for these exceptions became clear in about 1910, when Morgan, working on *Drosophila*, defined the notion of linkage. The idea of localising the elements carrying the genetic information, the genes or loci, linearly was put into practice in 1913 by Sturtevant, who made the first linkage map for five loci in *Drosophila*. Mathematical techniques required for the study of linkage and for mapping were first developed about 1935 by Fisher and Haldane (Bailey 1961). In 1980, a decrease in the number of studies concerned with the linkage mapping of plants was noticed (Tanksley and Rick 1980). Nevertheless, in more recent years, such studies have become numerous thanks to the development of the DNA restriction fragment length polymorphism (RFLP) technique. RFLP markers were first used in human genetics (Botstein et al. 1980; Drayna et al. 1984). Linkage maps of about 300 such markers were subsequently made for plant species such as maize (Helentjaris 1987), *Brassica* (Slocum et al. 1990) and potato (Gebhardt et al. 1991).

The construction of a genetic map begins with the estimation of recombination rates and with a test for linkage between all possible pairs of markers. The recombination rate θ between two loci is commonly estimated by maximum likelihood techniques. The resulting value is then tested for linkage using the null hypothesis $H_0: \theta = 0.5$, or equivalently the hypothesis that the loci are unlinked. The alternative hypothesis presumes that the loci are linked with a recombination rate equal to the estimator, and different from 0.5. The lod score ("log of the odds ratio" score) is one of the linkage tests most frequently used. For a given pair of loci, the method involves comparing the likelihood of linkage to that of independent segregation. The lod score is the decimal logarithm of the likelihood ratio. If this ratio is higher than a predetermined threshold, the null hypothesis will be rejected and the loci will be considered to be linked. In classical statistical tests, the decision of rejecting the null hypothesis depends directly on the choice of a first-type error level, which is the probability to reject the null hypothesis when it is true. The use of a critical value for the lod score cannot be justified in the same manner because it is not connected with a first-type error. A threshold of 3 is used in many studies, as suggested by Morton (1955), but the corresponding first-type error is usually not given. The aim of the present contribution was to compare two different tests, the lod score and a χ^2 test, in a linkage study made on maritime pine (*Pinus pinaster* Ait.).

Communicated by P. M. A. Tigerstedt

S. Gerber (✉)
ESV, bât 362, F-91405 Orsay Cedex, France

F. Rodolphe
INRA, Laboratoire de Biométrie, F-78370 Jouy en Josas, France

Materials and methods

Linkage data

The linkage data were obtained on 18 pine trees as described in a previous paper (Gerber et al. 1993). Two-dimensional polyacrylamide-gel electrophoresis of total proteins was performed on haploid megagametophytes (genetically equivalent to female gametes) of the seeds of this gymnosperm. An average of 12 megagametophytes per tree were individually analysed with this technique and the gels obtained were compared. The segregation of 84 loci affecting protein phenotypes (presence/absence, position or quantity modifications) were described.

Likelihood

To estimate linkage between two loci, only data coming from informative trees, that is heterozygous for both loci, can be used. Let m be the number of such trees for a given pair of loci. Let p be the probability for a tree to be in a given phase (coupling or repulsion). Let θ be the recombination rate. The probability $P_i(N_i, n_i, \theta)$ to observe n_i gametes in a given phase among the N_i gametes analysed for the ith tree is then:

$$P_i(N_i, n_i, \theta) = C_N^{n_i}[p(\theta^{n_i}(1-\theta)^{N_i-n_i}) + (1-p)(\theta^{N_i-n_i}(1-\theta)^{n_i})].$$

The likelihood for m independent trees is:

$$L = \prod_{i=1}^{m} P_i(N_i, n_i, \theta).$$

The log-likelihood can be written:

$$\mathrm{Log}(L) \approx \sum_{i=1}^{m} \log[p(\theta^{n_i}(1-\theta)^{N_i-n_i} + (1-p)(\theta^{N_i-n_i}(1-\theta)^{n_i})].$$

The values of $\theta (0 \leqq \theta \leqq 0.5)$ and of $p (0 \leqq p \leqq 1)$ that maximise this expression are examined. We are only interested in the θ value but p has to be included.

Lod scores

The lod scores were calculated with the Mapmaker computer package (Lander et al. 1987) using the human data type option as described elsewhere (Gerber et al. 1993).

χ^2 test

For a given pair of loci, N_i gametes of the ith tree heterozygous for both loci are observed. Among them, n_i are in a given phase. Under the null hypothesis (the loci are independent, $\theta = 0.5$), n_i has a binomial distribution with parameters N_i and 0.5. The S_i statistic:

$$S_i = \frac{(2n_i - N_i)^2}{N_i}$$

has a χ^2 distribution with one degree of freedom. It is not known which of the n_i or the $(N_i - n_i)$ gametes correspond to the parental (or to the recombinant) phase. However, the S_i statistic is invariant when n_i is replaced by $(N_i - n_i)$. A knowledge of the parental phase is therefore not needed to test for linkage.

A total of m trees are heterozygous for both loci. These trees are independent, so the S statistic:

$$S = \sum_{i=1}^{m} S_i$$

has a χ^2 distribution with m degrees of freedom, under the null hypothesis.

The S value and the probability of having observed a larger value by chance only, if the null hypothesis is true (which is the level of significance of the test), can be calculated for all informative pairs of loci. This was done using programs written with the Splus language (Becker et al. 1988). The null hypothesis is rejected when the significance level is smaller than a predetermined value, the first-type error.

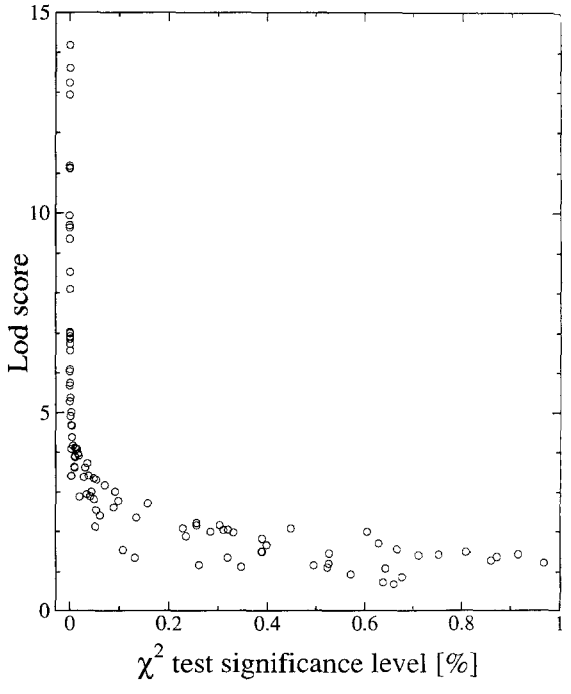
Results

With 84 loci, 3486 (84*83/2) combinations of markers by pairs were possible. Only 2711 pairs (78%) were found informative among the gametes of at least one tree in our linkage data. For each of these pairs the lod score value and the level of significance of the χ^2 test were calculated. Among the 2711 pairs, 129 had a χ^2 significance level smaller than 1%. The relationship between the lod score value and the χ^2 significance level for these pairs is given in Fig. 1. The lod score appears to be close to a decreasing function of the χ^2 significance level. The two tests are almost equivalent but not exactly: they have to be compared.

If a first-type error of 1% were to be judged sufficient to reject the null hypothesis with the χ^2 test, 129 pairs of loci would be declared linked. Among these pairs, 60 have a lod score smaller than 3 and 40 have a lod score smaller than 2. The lod score corresponds to a much more severe χ^2 test:

– 89 pairs of loci have a lod score greater than 2 for a maximal χ^2 significance level of 0.45%

Fig. 1 Relation between lod score values and χ^2 test significance levels



– 69 pairs have a lod score greater than 3 for a maximum significance level of 0.08%
– 48 pairs have a lod score greater than 4 for a maximum significance level of 0.015%

The two tests are equivalent for extreme values, i.e., strong linkage or total independence between loci. But near the values where thresholds are chosen, the tests are not similar. If, for instance, a first-type error of 0.08% is taken for the χ^2 test, pairs of loci with lod scores smaller than 3 will be grouped together. In contrast, high thresholds for the lod score warrant very low first-type errors.

The lod score is also very "cautious" for recombination rates (Fig. 2). With a minimal lod score of 3, the maximal recombination rate observed between loci declared linked is 0.25. With significance levels of 1% or 0.1% for the χ^2 test, higher recombination rates, close to 0.30, would result (Fig. 3).

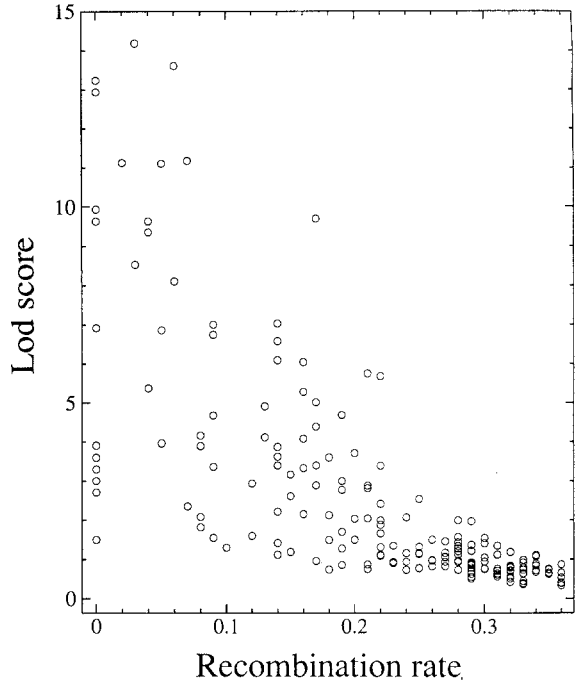
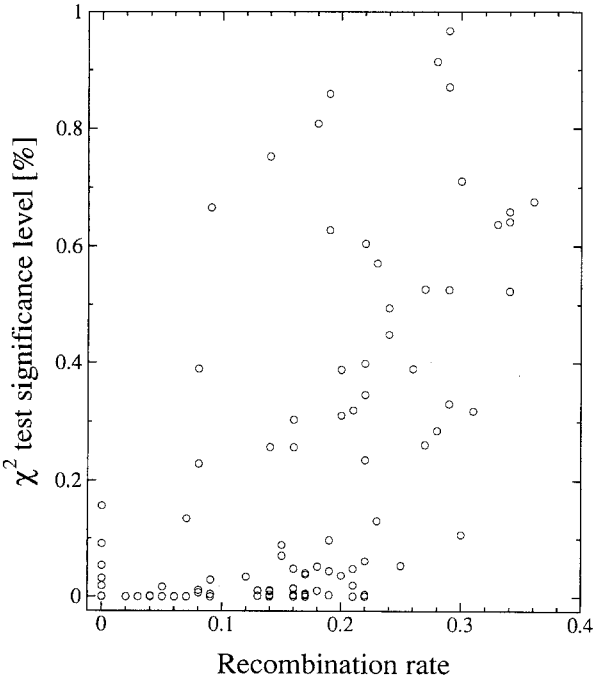


Fig. 2 Relation between lod score values and recombination rates

Fig. 3 Relation between χ^2 test significance levels and recombination rates



Discussion

The lod score method used in linkage studies usually refers to the paper of Morton (1955). However, there are differences between the current "lod score" and the "sequential test of likelihood ratio" suggested for Human genetics by this author. In this test, a θ_1 value is chosen at first for the recombination rate. The lod scores of the different families studied are calculated with this value and are cumulated. Two boundaries, A and B, are determined as functions of first- and second-type errors. If the sum of the lod scores is smaller than $\log_{10}(B)$ it will be concluded that the true θ value is greater than θ_1. If the sum is greater than $\log_{10}(A)$, the true θ value will be supposed smaller than 0.5 and the loci will be declared linked. If the sum is between $\log_{10}(A)$ and $\log_{10}(B)$ there will be no conclusion. When new data are collected, the sum is completed and when its value exceeds one of the boundaries, the test is concluded. In this context there are simple approximate relations between A, B and the first- and second-type errors (Wald 1947 cited by Morton 1955). With these relations, if a first-type error of 0.1% and a second-type error of 1% are requested, then $A \approx 1000$ and $B \approx 0.01$. The upper boundary is therefore $\log_{10}(A) = 3$. The value of 3, which is currently often used, appears here as initially suggested by Morton (1955).

In the present expression of the lod score, θ_1 is replaced by the maximum likelihood estimator of the recombination rate. Moreover, the sequentiality has disappeared and a global lod score is calculated on a sample of fixed size. With these modifications, the approximate relations between A, B and the first- and second-type errors no longer apply. Consequently, the actual values of the first- and second-type errors are smaller than that of the original test, according to Chotai (1984).

To find a more precise justification for the critical values chosen for the lod score with its modifications, Chotai (1984) refers to a result obtained by Haldane and Smith in 1947 for their likelihood ratio test with a fixed sample size. These authors showed that the likelihood ratio exceeds a value of A with a probability smaller than 1/A. According to Chotai (1984), this inequality roughly holds in the context of linkage analysis. More-

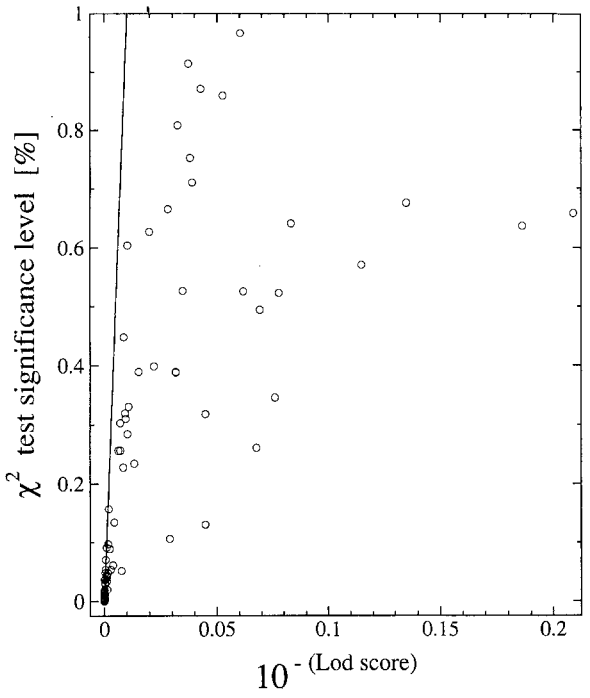


Fig. 4 Relation between χ^2 test significance levels and $10^{-\text{lod.score}}$ The line represents $(10^{-\text{lod.score}}) = (\chi^2$ test significance levels)

over, in certain cases, the first-type error associated with a lod score threshold of $\log_{10}(A)$ is nearly equal to 1/A. We compared the significance level of our χ^2 test with $1/A = 1/10^{-\text{lod score}}$ (Fig. 4). The line equating the significance level and 1/A is always greater than the observed values. For the highest lod scores, 1/A is rather close to the first-type error calculated, but when the lod score has lower values, the error is constantly smaller than 1/A. The test is very conservative. A critical value of 3 will correspond, in our case, to a maximum first-type error of $10^{-3} = 0.1\%$.

The geneticist will prefer no map to a false map: a severe test will limit the occurrence of a false positive when linkage is detected between independent loci. Moreover, with the increasing number of markers available, linkage studies have to deal with 100, 200 or 300 loci; that is, a maximum of 4950, 19900 or 44850 comparisons. The first-type error has to take these large numbers into account. For instance, calculating the lod scores of 81300 pairs of loci, Donis-Keller et al. (1987) required a minimum lod score of 4 to declare a linkage. With this threshold, they found all the same three linked pairs of loci corresponding to first-type errors, these loci being situated on different chromosomes.

The lod score technique was first developed for human genetics, where actual errors are smaller than expected. Plant geneticists work with simpler models, thanks to experimental populations: so conclusions may be different. The thresholds chosen, without being justified by a first-type error level, are diverse. For F_2 populations Messeguer et al. (1991) worked with thresholds between 2 and 3 whereas Devos et al. (1992) set them between 2 and 2.5. On recombinant inbred lines, Reiter et al. (1992) required threholds of 6. On the same kind of populations Ellis et al. (1992) compared the lod score with a χ^2 test. A threshold of 1.5 corresponds to a first-type error of 5%, whereas a threshold of 3 is associated with a significance level of 1% in their experiment. The significance level associated with a particular critical value of the lod score will obviously depend on the type of data used. It has to be estimated each time and no general rules can be given.

If the lod score may sometimes be too conservative for the detection of linkage, its use becomes more critical when characters in which the mode of inheritance are not precisely known are studied. For instance, in human genetics, when the linkage between genetic markers and complex diseases is investigated, the significance of the usual thresholds for lod scores can be difficult to assess. First-type errors may not be negligible in these cases (Clerget-Darpoux et al. 1990; Risch 1992). The study of quantitative trait loci in plant species could meet with the same kind of difficulty.

References

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

Becker RA, Chambers JM, Wilks AR (1988) The new S language. A programming environment for data analysis and graphics. Wadsworth & Brooks/Cole, California

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Chotai J (1984) On the lod score method in linkage analysis. Ann Hum Genet 48:359–378

Clerget-Darpoux F, Babron M-C, Bonaiti-Pellie C (1990) Assessing the effect of multiple linkage tests in complex diseases. Genet Epidemiol 7:245–253

Devos KM, Atkinson MD, Chinoy CN, Liu CJ, Gale MD (1992) RFLP-based genetic map of the homeologous group 3 chromosomes of wheat and rye. Theor Appl Genet 83:931–939

Donis-Keller H et al. (1987) A genetic linkage map of the human genome. Cell 51:319–337

Drayna D, Davies K, Hartley D, Mandel J-L, Camerino G, Williamson R, White R (1984) Genetic mapping of the human X chromosome by using restriction fragment length polymorphisms. Proc Natl Acad Sci USA 81:2836–2839

Ellis THN, Turner L, Hellens RP, Lee D, Harker CL, Enard C, Domoney C, Davies DR (1992) Linkage maps in pea. Genetics 130:649–663

Gebhardt C, Ritter E, Barone A, Debener T, Walkemeier B, Schachtschabel U, Kaufmann H, Thompson RD, Bonierbale MW, Ganal MW, Tanksley SD, Salamini F (1991) RFLP maps of potato and their alignment with the homeologous tomato genome. Theor Appl Genet 83:49–57

Gerber S, Rodolphe F, Bahrman N, Baradat Ph (1993) Seed-protein variation in maritime pine (*Pinus pinaster* Ait.) revealed by two-dimensional electrophoresis: genetic determinism and construction of a linkage map. Theor Appl Genet 85:521–528

Helentjaris T (1987) A genetic linkage map for maize based on RFLPs. Trends Genet 3:217–221

Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L (1987) Mapmaker: an interactive computer package for constructing primary genetic linkage maps for experimental and natural populations. Genomics 1:174–181

Messeguer R, Ganal M, de Vicente MC, Yound ND, Bolkan H, Tanksley SD (1991) High-resolution RFLP map around the root

knot nematode resistance gene (Mi) in tomato. Theor Appl Genet 82:529–536

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

Reiter RS, Williams JGK, Feldmann KA, Rafalski JA, Tingey SV, Scolnik PA (1992) Global and local genome mapping in *Arabidopsis thaliana* by using recombinant inbred lines and random amplified polymorphic DNAs. Proc Natl Acad Sci USA 89:1477–1481

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

Slocum MK, Figdore SS, Kennard WC, Suzuki JY, Osborn TC (1990) Linkage arrangement of restriction fragment length polymorphism loci in *Brassica oleracea*. Theor Appl Genet 80: 57–64

Tanksley SD, Rick CM (1980) Isozymic gene linkage map of the tomato: applications in genetics and breeding. Theor Appl Genet 57:161–170