Heredity **84** (2000) 437–440

Received 9 August 1999, accepted 15 November 1999

Optimal marker density for interval mapping in a backcross population

HANS-PETER PIEPHO*

Institut für Nutzpflanzenkunde, Universität Kassel, Steinstrasse 13, 37213 Witzenhausen, Germany

An important question in QTL mapping is the optimal choice of marker density. Using analytical results, it is shown for the case of interval mapping in a backcross population, that the power of QTL detection and the standard errors of genetic effect estimates are little affected by an increase of marker density beyond 10 cM. This finding confirms published simulation results by other authors.

Keywords: interval mapping, marker density, noncentrality parameter, power.

Introduction

An important question in QTL mapping is that of the optimal spacing of markers. Darvasi *et al.* (1993) performed a simulation study on the effect of marker density on the power to detect QTL by interval mapping (IM; Lander & Botstein, 1989) in a backcross population. They concluded with respect to the power or standard error of the estimate of gene effect that reducing marker spacing below 10 cM or 20 cM does not provide additional gains, regardless of the population size and gene effect. In this paper, I will use explicit formulae by Rebaï *et al.* (1995) to study the effect of marker spacing. The results will be compared to the simulation study by Darvasi *et al.* (1993).

Theory

If not indicated otherwise, the formulae presented in this section are taken from Rebaï *et al.* (1995). Consider an interval bordered by two codominant markers A and B with two alleles each (indexed 1 and 2) and a QTL with alleles Q_1 and Q_2. Two homozygous lines $A_1A_1Q_1Q_1B_1B_1$ and $A_2A_2Q_2Q_2B_2B_2$ are crossed. The F_1 hybrid is then backcrossed to $A_1A_1Q_1Q_1B_1B_1$, yielding a BC_1 population. The expected values for QTL genotypes Q_1Q_1 and Q_1Q_2 appearing in BC_1 are expressed as $\mu + a$ and $\mu - a$, respectively. We assume absence of double crossovers between the flanking markers. Haldane's mapping function is used throughout. Thus, the model for the BC_1 population is

$$Y_i = \mu + g_i a + e_i, \tag{1}$$

where Y_i is the phenotypic value of the ith individual ($i = 1, \ldots, n$), g_i is a dummy variable with $g_i = 1$ if the individual is Q_1Q_1 and $g_i = -1$, if it is Q_1Q_2, and e_i is a random normal error with variance σ^2. Without loss of generality, we will assume $\sigma^2 = 1$.

The score statistic for testing the null hypothesis of no QTL effect at a given position on the chromosome is

$$T(x) = \hat{a}(x)/\sqrt{\mathrm{var}[\hat{a}(x)]}, \tag{2}$$

where $\hat{a}(x)$ is the maximum likelihood estimator of a for given x and $\mathrm{var}[\hat{a}(x)]$ its asymptotic variance, and x is the recombination fraction between the left flanking marker A and the putative QTL position. $T(x)$ is distributed as N(0, 1) under H_0: $a = 0$ for a given putative QTL position x. $S(x) = [T(x)]^2$ is asymptotically equivalent to a likelihood-ratio (LR) test statistic for H_0 and conditional on x is distributed as a central χ^2 with one degree of freedom. In QTL mapping, the putative QTL position is not known, so the chromosome is scanned, performing multiple LR-tests. Thus, the $(1 - \alpha)$ quantile of χ^2_1 is not an appropriate threshold for controlling the chromosome-wise Type I error rate at α. Using results of Davies (1977, 1987), Rebaï *et al.* (1994, 1995) provided formulae for computing approximate critical thresholds C for controlling α, i.e.

$$\Pr[S(x) > C | a = 0] \leq \alpha. \tag{3}$$

The threshold C is found by solving numerically the equation (see Rebaï *et al.*, 1994):

$$\alpha = \Phi(\quad\sqrt{C}) + 2\exp\quad\tfrac{1}{2}C)\sum_{i=1}^{m}\arctan\left(\sqrt{\frac{p_i}{1\quad p_i}}\right), \tag{4}$$

*Correspondence. E-mail: piepho@wiz.uni-kassel.de

© 2000 The Genetical Society of Great Britain.

where $\Phi(.)$ is the cumulative standard normal distribution function, p_i is the recombination fraction between markers flanking the ith interval, and m is the number of intervals on the chromosome. The approximate thresholds were shown to be very close to thresholds based on simulation for marker spacing up to 5 cM (Rebaï *et al.*, 1994). The value of C depends on the length of the chromosome and on the spacing of the markers. We will use these thresholds in power computations.

To assess the power of QTL detection, we will consider the probability that $S(x)$ exceeds C at the true QTL position x_0 for a given value of the genetic effect a, i.e.

$$\text{Power} = \Pr[S(x_0) > C|a, x_0]. \tag{5}$$

This will provide a lower bound for the power of the procedure (Davies, 1977; Rebaï *et al.*, 1995). The power in eqn (5) can be computed by noting that under the alternative hypothesis H_1: $a \neq 0$, $S(x_0)$ follows a noncentral χ^2 distribution with one degree of freedom and noncentrality parameter

$$\lambda = na^2\{1 - 4x_0(p - x_0)/p\}, \tag{6}$$

where n is the sample size and p is the recombination fraction between the flanking markers A and B. The power depends on x_0, the QTL position, through the noncentrality parameter.

Finally, given a putative QTL is located at position x_0, the conditional large-sample variance of $\hat{a}(x)$ (i.e. for large n) under H_0: $a = 0$ is

$$\text{var}[\hat{a}(x_0)] = p/[n(p - 4x_0p + 4x_0^2)]. \tag{7}$$

This conditional variance provides a lower bound for the unconditional variance var$(\hat{a})$, i.e. the variance for unknown x_0, for which no simple analytical result could be obtained. Also, no simple expressions are available for the conditional or the unconditional variance under the alternative H_1: $a \neq 0$. Thus, simulations as performed by Darvasi *et al.* (1993) are necessary to study these variances.

Results

First consider the asymptotic variance of $\hat{a}(x_0)$ under H_0 in eqn (7). Because the putative QTL is flanked by markers A and B, x_0 will lie between 0 and p. Thus, we may write

$$x_0 = \varepsilon p, \tag{8}$$

where $0 \leq \varepsilon \leq 1$. Using eqn (7), the asymptotic variance can be expressed as

$$\text{var}[\hat{a}(x_0)] = 1/\{n[1 - 4\varepsilon(1 - \varepsilon)p]\}. \tag{9}$$

Equation (9) shows two interesting facts. First, for a given marker spacing p, the variance is maximal when the putative QTL is located halfway between the flanking markers, i.e. when $\varepsilon = 0.5$. The maximal value is $\text{var}[\hat{a}(x_0)] = n^{-1}(1 - p)$. The minimum variance is achieved when the QTL is at one of the markers, i.e. $\varepsilon = 0$ or $\varepsilon = 1$, whence $\text{var}[\hat{a}(x_0)] = n^{-1}$. Secondly, for any relative position of the QTL (x_0), the variance is monotonically decreasing in p, reaching its maximum for an infinitely dense marker spacing, i.e. when $p \rightarrow 0$. The limiting variance is $\lim_{p \rightarrow 0} \text{var}[\hat{a}(x_0)] = n^{-1}$. Thus, the variance cannot be reduced below the limit n^{-1}, regardless of the marker spacing.

An important question is the magnitude of shift in $\text{SE}[\hat{a}(x_0)] = \sqrt{\text{var}[\hat{a}(x_0)]}$ as we change the marker spacing p. The standard error depends on the relative position of the putative QTL, which we indicate by the notation SE$[\hat{a}(x_0),\varepsilon]$. For a simple analysis, we average SE$[\hat{a}(x_0),\varepsilon]$ across the interval $0 \leq \varepsilon \leq 1$, giving equal weight to each value of ε. This is equivalent to assigning a uniform prior to ε on the interval [0, 1]. We find by straightforward calculations that

$$\begin{aligned}\overline{SE[\hat{a}(x_0)]} &= \int_0^1 \text{SE}[\hat{a}(x_0), \varepsilon]\text{d}\varepsilon \\ &= \frac{1}{\sqrt{4pn}}\left[\log\left(\frac{1}{\sqrt{4p}} + \frac{1}{2}\right) - \log\left(\frac{1}{\sqrt{4p}} - \frac{1}{2}\right)\right].\end{aligned} \tag{10}$$

Figure 1 shows a plot of $\overline{\text{SE}[\hat{a}(x_0)]}$ vs. d, the distance between flanking markers in cM based on Haldane's mapping function, for $n = 1$. The change of $\overline{\text{SE}[\hat{a}(x_0)]}$ is not dramatic for d between 0 and 10 cM. Thus, an increase in marker density beyond 10 cM does not do much to improve the accuracy of the QTL effect estimate.

We also computed the average of the power $\Pr[S(x_0) > C|a, x_0)]$ across the interval $[0, p]$. A chromosome length of 100 cM was assumed. The critical value C in eqn (4) was computed for $\alpha = 5\%$ and equidistant marker spacing. Because explicit integration was not feasible, we computed the average across a uniform grid of 100 steps for x_0, i.e. for $x_0 = p\varepsilon$ with ε increasing from 0 to 1 in steps of 0.01. This is the same approach as that used by Rebaï *et al.* (1995). Figure 2 shows the average power of $S(x_0)$ for different values of the genetic effect a and a sample size of $n = 200$. For small genetic effects ($a = 0.1$ and $a = 0.2$), the power increases with

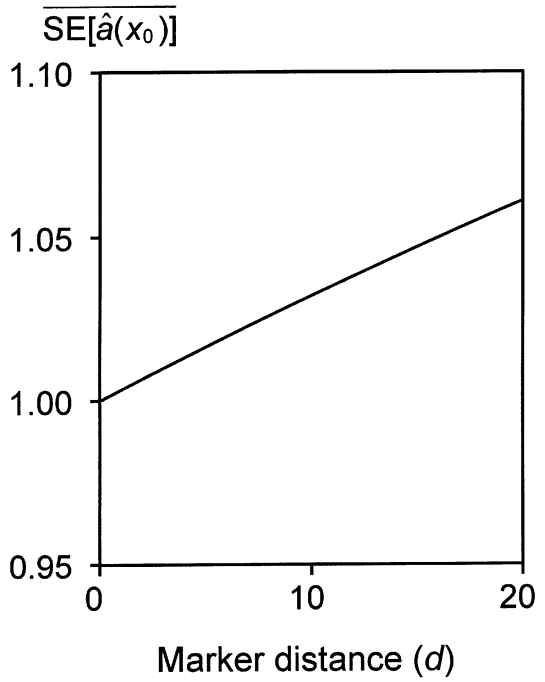

Fig. 1 Standard error of maximum likelihood estimate of a for given true QTL position x_0 under H_0: $a = 0$ as a function of marker distance d (markers assumed to be equally spaced across the chromosome) and relative QTL position between flanking markers (ε). Standard error averaged across flanking marker interval.

increasing marker spacing, whereas for larger effects ($a = 0.25$ and $a = 0.3$) the power is rather stable between spacings of 0 and 20 cM. This result is surprising at first sight, because the noncentrality parameter λ increases linearly with decreasing p, so one would expect an increase of power as spacing gets denser. The linearity of λ in p can be demonstrated by expressing x_0 as $x_0 = p\varepsilon$ which leads to

$$\lambda = na^2[1 \quad 4\varepsilon(1 \quad \varepsilon)p]. \tag{11}$$

Thus, as we decrease p, the probability distribution function (p.d.f.) of $S(x_0)$ is shifted towards larger values of $S(x_0)$ under the alternative. This does not necessarily result in a gain of power, however, because at the same time the critical threshold C needs to be increased in order to control the chromosome-wise Type I error rate. An increase in power results only if the effect of shifting the p.d.f. of $S(x_0)$ under the alternative more than offsets the effect of increasing C. The amount of shift in the p.d.f. of $S(x_0)$ will be larger for large effect sizes a, as λ is quadratic in a. This explains why increasing the marker density, i.e. decreasing p, tends not to pay off for smaller values of a. Also note that the noncentrality parameter cannot be increased beyond na^2, the limit as $p \to 0$.

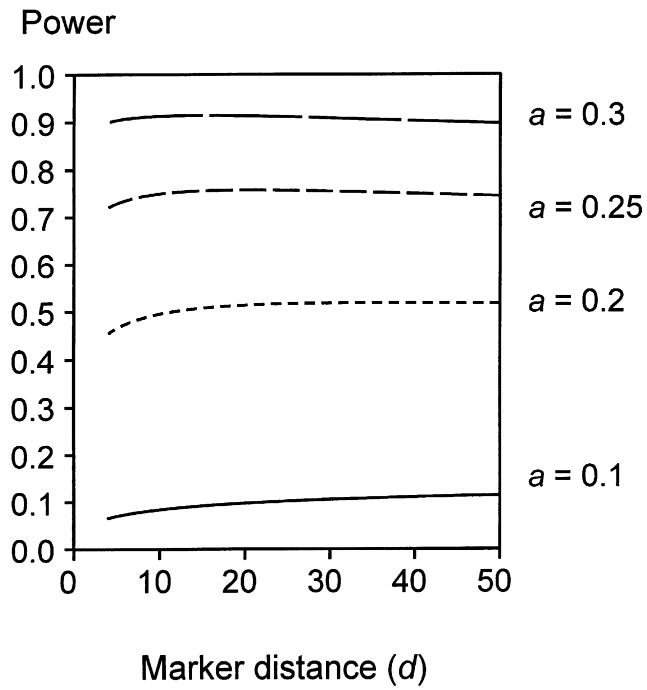

Fig. 2 Average power of QTL detection at true QTL position x_0 for different effect sizes a and different marker distances d (markers assumed to be equally spaced across the chromosome). Sample size: $n = 200$. Chromosome length: 100 cM. Power is averaged across a grid for $x_0 = p\varepsilon$ where p is the recombination fraction and ε ranges from 0 to 1 with a step size of 0.01.

The power curves will have the same shape for any value of n, except that for a given effect size a the curves will be moved upwards. The reason for this is that for constant values of p and x_0, the noncentrality parameter λ is proportional to na^2. For example, the power curve for $n = 800$ and $a = 0.2$ will be the same as that for $n = 200$ and $a = 0.4$, because $800 \times 0.2^2 = 200 \times 0.4^2 = 32$. The general conclusion to be drawn from Fig. 2 is that for detecting large QTL effects, a more dense marker spacing is preferable, whereas for detecting small QTLs, a less dense spacing is better. A similar conclusion was obtained by Darvasi *et al.* (1993) based on the empirical width of confidence intervals for QTL location. It is impossible to achieve optimal power for all effect sizes with a single spacing. Because the power for detecting large QTLs will generally be larger than for small QTLs, one might consider choosing the spacing so that the power for smaller QTLs is nearly optimal, i.e. to choose a less dense spacing. In any case the dependence of power on marker spacing is not large, so there seems little advantage in choosing a very fine marker spacing. I also used the formula by Dupuis & Siegmund (1999) for approximate power and obtained very similar results (not shown).

Final remarks

All conclusions in the preceding section agree well with the simulation results by Darvasi *et al.* (1993), though it must be kept in mind that I have used the conditional variance (under H_0) and power, which only provide bounds on the unconditional variance (under H_0) and power. I investigated only the conditional variance under H_0, because no simple analytical expressions are available for the unconditional variance and for the alternative H_1; simulations as presented in Darvasi *et al.* (1993) are the most useful means to study these latter variances. The advantage of using simple analytical expressions, if available, is that computationally demanding simulations can be avoided, that a deeper insight into the causes of power differences, etc. can be gained and that more general conclusions can be drawn. This study confirms earlier simulation-based findings that marker spacing has only limited influence on the power to detect a QTL. Thus, the design can be optimized for other purposes such as marker-assisted selection and transfer of target genes (Frisch *et al.*, 1999).

This paper has considered the effect of marker spacing in IM for a backcross population, because analytical results are available for this case. It is conjectured that the general conclusions would not be grossly different for other populations such as F_2 and for composite interval mapping (CIM) (Jansen, 1993; Zeng, 1993). I have not studied the effect on the width of confidence intervals for QTL position, because simple analytical results are not available for the case of an intermediate map density (>1 cM) (Mangin *et al.*, 1994; Visscher *et al.*, 1996; Mangin & Goffinet, 1997; Dupuis & Siegmund, 1999). Simulations (Darvasi *et al.*, 1993; Visscher *et al.*, 1996) indicate that the expected width changes only marginally with marker spacing, whereas it decreases notably as the genetic effect size *a* increases.

Acknowledgements

This paper was written while the author was visiting the Department of Biometrics, College of Agriculture and Life Sciences, Cornell University, Ithaca, New York, USA. Support of the Heisenberg-Programm of the Deutsche Forschungsgemeinschaft (DFG) is thankfully acknowledged. I would like to thank Hugh Gauch, Jr and two anonymous referees for helpful comments on an earlier draft.

References

DARVASI, A., WEINREB, A., MINKE, V., WELLER, J. I. AND SOLLER, M. 1993. Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics*, **134**, 943–951.

DAVIES, R. B. 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **64**, 247–254.

DAVIES, R. B. 1987. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**, 33–43.

DUPUIS, J. AND SIEGMUND, D. 1999. Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics*, **151**, 373–386.

FRISCH, M., BOHN, M. AND MELCHINGER, A. E. 1999. Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. *Crop Sci.*, **39**, 967–975.

JANSEN, R. C. 1993. Interval mapping of multiple quantitative trait loci. *Genetics*, **135**, 205–211.

LANDER, E. S. AND BOTSTEIN, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

MANGIN, B. AND GOFFINET, B. 1997. Comparison of several confidence intervals for QTL detection. *Heredity*, **78**, 345–353.

MANGIN, B., GOFFINET, B. AND REBAÏ, A. 1994. Constructing confidence intervals for QTL location. *Genetics*, **138**, 1301–1308.

REBAÏ, A., GOFFINET, B. AND MANGIN, B. 1994. Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138**, 235–240.

REBAÏ, A., GOFFINET, B. AND MANGIN, B. 1995. Comparing power of different methods for QTL detection. *Biometrics*, **51**, 87–99.

VISSCHER, P. M., THOMPSON, R. AND HALEY, C. S. 1996. Confidence intervals in QTL mapping by bootstrapping. *Genetics*, **143**, 1013–1020.

ZENG, Z. B. 1993. Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 10972–10976.