Theor Appl Genet (1989) 77:815–819

THEORETICAL AND APPLIED GENETICS

Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines

S. P. Simpson

AFRC Institute of Animal Physiology and Genetics Research, Edinburgh Research Station, West Mains Road, Edinburgh EH9 3JQ, UK

Received November 1, 1988; Accepted February 15, 1989
Communicated by L. D. Van Vleck

Summary. In segregating populations, large numbers of individuals are needed to detect linkage between markers, such as restriction fragment length polymorphisms (RFLPs), and quantitative trait loci (QTL), limiting the potential use of such markers for detecting linkage. Fewer individuals from inbred lines are needed to detect linkage. Simulation data were used to test the utility of two methods to detect linkage: maximum likelihood and comparison of marker genotype means. When there is tight linkage, the two methods have similar power, but when there is loose linkage, maximum likelihood is much more powerful. Once inbred lines have been established, they can be screened rapidly to detect QTL for several traits simultaneously. If there is sufficient coverage of the genome with RFLPs, several QTL for each trait may be detected.

Key words: Quantitative traits – Restriction fragment length polymorphisms – Inbred lines

Introduction

Many quantitative traits of economic importance are likely to be under the control of several genes, each with a relatively small effect. These traits are often exploited using direct selection. Measurement of some traits may be expensive (Neinhuis et al. 1987) or inconvenient, for example, sex-limited traits (Smith and Simpson 1986). In these cases, indirect selection using marker loci may be appropriate. However, before indirect selection can be exploited, we need to be able to detect linkage between marker loci and the loci controlling the quantitative trait (QTL – quantitative trait loci). In segregating populations, large numbers of individuals, probably several thousand, are needed to detect linkage (Soller and Beckmann 1983), especially when alleles are at extreme frequencies or if the marker locus exhibits dominance. The use of restriction fragment length polymorphisms (RFLPs) as markers, which have the advantage of being numerous and codominant, has been considered extensively in segregating populations (e.g. Beckmann and Soller 1983). However, extreme allele frequencies can limit their potential usefulness. Bennett et al. (1982) have used recombinant inbred lines to map RFLPs and an extension of this would be to use inbred lines to detect linkage between RFLPs and QTL.

Ellis (1986) considered the possibility of detecting linkage using inbred lines, and concluded that "there are severe limitations to the utility of this approach". Despite this, Tanksley et al. (1982), using 12 enzymatic loci in tomatoes, were able to detect 21 QTL using an interspecific backcross, a potentially less powerful approach. This paper evaluates two different appoaches which can be used to detect linkage between QTL and RFLPs in inbred lines: comparison of marker genotype means, which was considered by Ellis (1986), and maximum likelihood, which has the advantage of also allowing estimation of recombination rates and the size of the effects of the QTL. Although Ellis (1986) only considered selfing in plants, both selfing and brother-sister inbreeding are considered here.

Materials and methods

Origins of the data

The data are assumed to have arisen in the following way. Initially, there were two distinct populations, possibly inbred lines, which are assumed to be homozygous at many different loci. In particular, they are assumed to be homozygous at a marker locus with genotype $P_1 P_1$ or $P_2 P_2$ and a QTL with genotype

C_1C_1 or C_2C_2. These two populations are crossed to form a heterozygous population where all individuals have the genotype $P_1P_2C_1C_2$. This population forms the basis for a set of new distinct inbred lines, either by selfing in plants or brother-sister mating in animals. The effect of inbreeding on a heterozygous population is to reduce heterozygosity, and eventually each line becomes homozygous. The rate at which homozygosity is approached can be calculated using the formulae derived by Haldane and Waddington (1931). Due to recombination, in the P_1P_1 class, a proportion $(1-R)$ will have the genotype $P_1P_1C_1C_1$ and a proportion (R) will have the genotype $P_1P_1C_2C_2$, whereas in the P_2P_2 class a proportion $(1-R)$ will have the genotype $P_2P_2C_2C_2$ and a proportion (R) will have the genotype $P_2P_2C_1C_1$. The proportion R is a function of the recombination rate r, and is $2r/(1+2r)$ in selfing populations and $4r/(1+6r)$ for brother-sister mating (Haldane and Waddington 1931). In both cases $R = 1/2$ when $r = 1/2$, and $R = 0$ when $r = 0$.

Distribution of QTL traits in inbred lines

If the quantitative trait is normally distributed with mean m_1 and variance s_1^2 when the genotype is C_1C_1 and mean m_2 and variance s_2^2 when the genotype is C_2C_2, the resulting new inbred lines will be a mixture of two normal distributions. The new inbred lines are assumed to be fully inbred. We will assume $s_1^2 = s_2^2$ and denote this value by s^2, and consider the symmetric case where there are N lines of each marker genotype. We will initially assume that observations are made on one individual from each inbred line to give a total of $2N$ observations. The variation about the mean is assumed to be in part environmental and in part due to other unlinked QTL. In practice, each line will consist of many individuals or replicates, and the effect of using observations on replicates within lines will be discussed more fully later.

Let the $2N$ observations be denoted by x_{ij}, where i denotes the genotype, P_1P_1 or P_2P_2, and j the replicates, or lines, within each marker genotype.

Likelihood ratio test for linkage

Given the data, we can form the likelihood which can be maximized to find maximum likelihood estimates of m_1, m_2, s and r, the recombination rate. For the inbred lines, the likelihood is

$$L = \prod_{j=1}^{N} \{1-R)\,\phi((x_{1j}-m_1)/s) + R\,\phi((x_{1j}-m_2)/s)\}\ x$$
$$= \prod_{j=1}^{N} \{R\,\phi((x_{2j}-m_1)/s) + (1-R)\,\phi((x_{2j}-m_2)/s)\}$$

where ϕ is the standardized normal density function. The substitutions $R = 2r/(1+2r)$ and $R = 4r/(1+6r)$ give the likelihoods for selfing and brother-sister mating, respectively.

The likelihood can be maximized with respect to the parameters m_1, m_2, s and r for the hypothesis of linkage and with respect to m_1, m_2 and s, with the constraint $r = 0.5$ for absence of linkage, i.e. free recombination. The natural logarithm of the ratio of the two maximized likelihoods, the likelihood ratio test statistic, is asymptotically distributed as $-1/2\,(\chi_1^2)$. The expected value of the test statistic or the number of inbred lines required to detect linkage cannot easily be derived explicitly, but can be estimated using Monte Carlo simulation.

Simulation

Data sets consisting of 200 observations, 100 inbred lines within each marker genotype P_1P_1 and P_2P_2, each comprised of a single individual, were simulated for various values of $m_1 - m_2$ and r. In the simulation we set $m_1 = 0$ and $s = 1$, as these are location and scale parameters which do not affect the size or power of the test. The results are presented in terms of $d = (m_1 - m_2)/s$, the standardized difference between the means.

The likelihood was maximized with respect to m_1, m_2 and s, first with r unconstrained $(0 \leq r \leq 0.5)$ and then with r fixed at $r = 0.5$ to obtain a χ_1^2 statistic. Even though m_1 and s were fixed, they were estimated, since in practice they would be unknown parameters which would need estimating. Estimation was by iterative maximization using the sub-routine GEMINI (Lalouel 1979). For each value of d and r chosen, 200 data sets were simulated. The proportion of observed likelihood ratio test statistics greater than the tabulated 5% value of 3.84 would give an estimate of the power of the test at the 5% significance level, but such a value would only be relevant to a sample size of 200 inbred lines. A useful statistic would be the number of inbred lines required to detect linkage at, say, the 5% significance level. When the hypothesis of linkage is true, the likelihood ratio test statistic follows a non-central $\chi_1^2(\lambda_N)$ distribution with non-centrality parameter $\lambda_N > 0$ (Lancaster 1969). For large samples the non-centrality parameter depends linearly on the sample size. This can be exploited to estimate the number of inbred lines needed to detect linkage at any given significance level. Furthermore, the mean of the observed likelihood ratio test statistics can be scaled to obtain estimates of the non-centrality parameter for different sample sizes and to calculate the power function.

Comparison of marker genotype means

The expected mean of the observations will be $(1-R)m_1 + Rm_2$ in the P_1P_1 lines and $Rm_1 + (1-R)m_2$ in the P_2P_2 lines. The expected difference between the observed means is $(1-2R)(m_1-m_2) = (1-2R)ds$, where d is the standardized difference between the QTL means. As the observed variance within each marker genotype (P_1P_1 and P_2P_2) is

$$s^2 + (m_1-m_2)^2 R(1-R) = s^2(1 + d^2 R(1-R)),$$

the variance of the difference between the means is

$$2s^2(1 + d^2R(1-R))/N.$$

From this we obtain a Student's t-statistic

$$t = \frac{(1-2R)ds}{\sqrt{2s^2(1+d^2R(1-R))/N}}.$$

When the marker locus and the QTL are linked, the sample size required to detect linkage at the 5% significance level (and 50% power) is

$$N > \frac{1.96^2 \cdot 2(1+d^2R(1-R))}{(1-2R)^2 d^2} \qquad (1)$$

The t-statistic can also be used to calculate the power of the test for various sample sizes. For ease of computation, the power has been calculated using a normal approximation to Student's t-distribution and, hence, for small sample sizes the power will be slightly under-estimated.

Ellis (1986) used a similar t-statistic

$$t = \frac{M - M(0.5)}{\sqrt{S^2 + S^2(0.5))/N}}$$

where M and S^2 are the observed mean and variance in the P_1P_1 class, and M (0.5) and S^2 (0.5) are the expected mean and variance when linkage is absent. Although not stated explicitly, in the absence of linkage the overall mean and variance will be estimates of M (0.5) and S^2 (0.5). Since the P_1P_1 class is used in both means and variances, this is not an efficient test statistic.

Results

Establishment of inbred lines

In selfing populations, very few generations of inbreeding are needed to obtain inbred lines. Many more generations of inbreeding may be needed with brother-sister mating. Table 1 gives the number of generations of inbreeding needed to obtain lines which are 90%, 99% and 99.9% homozygous with respect to a pair of loci, for various recombination rates.

Distribution of the likelihood ratio test statistic

When there is free recombination, the test statistic (-2 log likelihood ratio) should be asymptotically distributed as χ_1^2, the central chi-squared distribution. Data sets were simulated with $r = 0.5$ for various values of d. The mean of the test statistics should be one and 5% should be greater than 3.84. For $d > 1.0$, the simulations were consistent with this, but for smaller values of d the average test statistic was smaller than expected. This was due to convergence problems. The estimate of r was constrained to be positive. In some replicates, the iterative procedure converged to the boundary $r = 0.0$ and the global maximum was not attained. These replicates were not excluded, since if they were omitted the test statistic may have been over-estimated. As a consequence, the average test statistic tended to be under-estimated.

When the loci are linked, the test statistic is distributed as $\chi_1^2(\lambda_N)$ and the non-centrality parameter λ_N is a linear function of N. Linearity was confirmed by simulating data sets with sample sizes ranging from 50 to 400. Supplementary tests also confirmed that test statistic is distributed as non-central $\chi_1^2(\lambda_N)$. Again, convergence problems were experienced with small values of d, which resulted in the average test statistic being under-estimated.

Table 1. Number of generations of inbreeding to obtain various levels of homozygosity

Homozygosity	Selfing			Brother-sister mating		
	90.0%	99.0%	99.9%	90.0%	99.0%	99.9%
Recombination rate						
0.0	4	7	10	12	23	34
0.1	4	8	11	17	35	52
0.2	5	8	11	19	36	53
0.3	5	8	11	19	36	54
0.4	5	8	11	19	37	54
0.5	5	8	11	19	37	54

Number of inbred lines needed to detect linkage

The number of inbred lines needed to detect linkage using comparison of marker genotype means can be calculated using formula (1), but the power of the test is only 50%. For comparison, a likelihood ratio test of the same power is needed. The mean value of the test statistic can be equated to its expected value $1+\lambda_N$. Since λ_N depends linearly on the sample size, the value of N such that $1+\lambda_N > 3.84$ gives a test with power 50%. Since the test statistic is under-estimated for small d, the estimate of the number of lines needed will be conservative.

The total number of inbred lines needed to detect linkage at the 5% significance level are given in Table 2 for d ranging from 0.25 to 3.0 standard deviations (sd)

Table 2. Total number of inbred lines needed to detect linkage at 5% significance level

Standardized difference between means $d=(m_1-m_2)$		Recombination rate $-r$[a]							
		Selfing				Brother-sister mating			
		0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30
$d=0.25$	χ^2[b]	267	355	617	1,243	293	561	1,012	2,176
	t[c]	369	558	1,356	3,991	517	995	3,353	12,232
$d=0.50$	χ^2	77	112	173	424	105	163	384	968
	t	94	143	352	1,041	132	257	874	3,169
$d=0.75$	χ^2	33	56	116	271	49	86	240	503
	t	43	66	166	495	61	121	415	1,523
$d=1.00$	χ^2	19	31	79	189	28	55	151	344
	t	25	39	101	303	36	73	254	937
$d=2.00$	χ^2	7	11	26	85	10	20	68	199
	t	8	13	38	119	12	27	99	373
$d=3.00$	χ^2	4	7	18	56	7	13	45	131
	t	4	9	26	85	9	18	71	268

[a] Proportion recombinant is $R=2r/(1+2r)$ for selfing and $R=4r/(1+6r)$ for brother-sister mating

[b] χ^2 denotes likelihood ratio test statistic

[c] t denotes comparison of marker genotype means

Table 3. Power of a 5% test using a total of 200 inbred lines

Standardized difference between means $d=(m_1-m_2)$		Recombination rate $-r$[a]							
		Selfing				Brother-sister mating			
		0.05	0.10	0.20	0.30	0.05	0.10	0.20	0.30
$d=0.25$	χ^2[b]	0.42	0.37	0.28	0.23	0.40	0.29	0.24	0.20
	t[c]	0.30	0.21	0.11	0.06	0.23	0.14	0.07	0.04
$d=0.50$	χ^2	0.82	0.69	0.54	0.33	0.72	0.56	0.35	0.24
	t	0.82	0.64	0.31	0.13	0.67	0.41	0.15	0.07
$d=0.75$	χ^2	0.99	0.92	0.68	0.42	0.94	0.79	0.45	0.31
	t	0.99	0.92	0.58	0.24	0.94	0.71	0.28	0.11
$d=1.00$	χ^2	1.00	0.99	0.81	0.52	1.00	0.92	0.59	0.37
	t	1.00	0.99	0.79	0.36	1.00	0.90	0.41	0.15
$d=2.00$	χ^2	1.00	1.00	1.00	0.79	1.00	1.00	0.87	0.50
	t	1.00	1.00	0.99	0.72	1.00	1.00	0.79	0.30
$d=3.00$	χ^2	1.00	1.00	1.00	0.91	1.00	1.00	0.96	0.64
	t	1.00	1.00	1.00	0.85	1.00	1.00	0.91	0.40

[a] Proportion recombinant is $R=2r/(1+2r)$ for selfing and $R=4r/(1+6r)$ for brother-sister mating

[b] χ^2 denotes likelihood ratio test statistic

[c] t denotes comparison of marker genotype means

and R ranging from 0.09 to 0.43, corresponding to r ranging from 0.05 to 0.3. These were derived from the mean values of the likelihood ratio statistic. Recombination rates greater than 0.2 represent loose linkage. While they might allow chromosomal assignment and demonstrate the existence and effects of QTL, the markers are not likely to be of much use for marker-based selection. In practice, the most useful markers are likely to be those very closely linked to the QTL with recombination rates of 0.1 or less. Differences between the QTL homozygote means greater than 1.0 sd, and certainly greater than 2.0 sd, would indicate a major gene segregating in the population, whereas QTL are more generally considered to be genes which have smaller effects on the quantitative trait (0.2–1.0 sd).

When the marker locus and the QTL are tightly linked ($r<0.1$) slightly fewer lines are needed using likelihood ratio. As the recombination rate increases, the number of lines needed for the t-test increases very rapidly. The increase in the number of lines needed for the likelihood ratio test increases much more slowly, especially when the difference between QTL means is small. When there is loose linkage, about twice as many lines are needed under brother-sister mating compared with selfing, but the difference is less for tight linkage.

Size and power of the tests for fixed sample size

The size of the test is the probability of claiming that linkage exists when it does not, and the power is the probability of being able to detect linkage when it does exist. Table 3 gives the power of a 5% test using a total of 200 inbred lines. In general, the likelihood ratio test statistic is superior, since it exploits the bimodality or skewness of the data as well as the difference between the means. When there is tight linkage, the two methods are almost equally powerful, but for loose linkage the likelihood ratio approach is much more powerful.

Discussion

In practice, each line will be represented by several individuals, e.g. litter mates in inbred mice lines or several offspring from a single selfed plant. With continued inbreeding within each line, the individuals will become genetically identical. The effect of the environmental variation can be reduced by using the mean of individuals within lines. The variation between lines is then due to the QTL under consideration and the remaining QTL. This will increase the power of the test to detect linkage, but the amount by which it can be increased is limited by the amount of variation in the trait which is attributable to the remaining QTL and how much is environmental.

When the heritability of the quantitative trait is known, the total variation due to QTL is known, but not the amount of variation due to individual QTL. If the trait is highly heritable, the environmental component will be small and fewer individuals per line will be required to detect linkage. When the heritability is known, the number of individuals per line required can be estimated. In general, more information is gained from generating more lines than from increasing the number of replicates within a line. For example, for a heritability of 0.3, a new line consisting of 100 individuals may be more informative than adding 10,000 new individuals to existing lines.

A priori, the size of the contribution of each QTL will be unknown and the number of lines required will be unknown. If sufficient lines are available and the number of available RFLP markers is sufficiently large, we should be able to detect many QTL and be able to estimate how much of the variation is attributable to each QTL and how much remains unaccounted for. This can be achieved by resolving the components of variation using a hierarchical analysis of variance, or using maximum likelihood after reformulating the likelihood to include variation due to environment and the remaining QTL.

The prime objective is to detect QTL closely linked to the RFLPs, and by definition the standardized differences between means due to an individual QTL is taken to be relatively small. Under these circumstances, for example if $r \leq 0.2$ and $d = 1.00$, around 100 of each marker genotype should suffice. The inbreeding scheme should ensure approximately equal numbers of each marker genotype. When there is tight linkage, maximum likelihood is only slightly more powerful than a t-test or analysis of variance and is more laborious, as it requires iterative methods for estimation. It has the advantage, however, of enabling estimates of the size of the effect of the QTL and recombination rate to be calculated.

Large numbers of RFLPs can be generated. A suitable strategy to employ them would be to use a simple t-test or analysis of variance to screen inbred lines for markers which appear to be linked to QTL, and then use maximum likelihood to estimate the difference between the means and the recombination rate for only those markers.

The greatest problem is in the amount of effort needed to produce the inbred lines. However, using inbred lines, we should be able to detect more QTL with 100 inbred lines than using several thousand individuals from a segregating population. Although relatively large numbers of inbred lines are required to detect linkage and they are difficult to generate, once the inbred lines have been established they can be rapidly screened to detect QTL for several traits. If there is sufficient coverage of the genome by the RFLPs, several QTL for each trait may be detected. In plants and Drosophila, and possibly mice, the formation of many inbred lines is possible, but this would be too costly in large animal species. However, murine models could be used to detect QTL for growth or production traits, which could then be synthesized and used for further investigations (Nadeau and Taylor 1984). Ellis (1986) concluded that inbred lines are inappropriate for the detection of linkage. This reappraisal concludes, to the contrary, that when sufficient new inbred lines can be generated, they can be a powerful tool for the detection of QTL for several traits, all measured simultaneously. Furthermore, as there are approximately equal numbers of individuals with each RFLP allele in inbred populations, they are likely to be more informative than outcrossed populations where the frequencies of the RFLP alleles are less equal.

Acknowledgements. I am grateful to Drs. Ch. Smith and R. Thompson for invaluable discussions of this paper.

References

Beckmann JS, Soller M (1983) Restriction fragment length polymorphism in genetic improvement: methodologies, mapping and costs. Theor Appl Genet 67:35–43

Bennett KL, Lalley PA, Barth RK, Hastie ND (1982) Mapping the structural genes coding for major urinary proteins in the mouse: Combined use of recombinant inbred strains and somatic cell hybrids. Proc Natl Acad Sci USA 79:1220–1224

Ellis THN (1986) Restriction fragment length polymorphism markers in relation to quantitative characters. Theor Appl Genet 72:1–2

Haldane JBS, Waddington CH (1931) Inbreeding and linkage. Genetics 16:357–374

Lalouel J-M (1979) GEMINI – A computer program for optimization of general non-linear functions. Tech Report 14, University of Utah

Lancaster HO (1969) The Chi-squared distribution. Wiley, New York

Nadeau JH, Taylor BA (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. Proc Natl Acad Sci USA 81:814–818

Neinhuis J, Helentjaris T, Slocum M, Ruggero B, Schaefer A (1987) Restriction fragment length polymorphism analysis of loci associated with insect resistance in tomato. Crop Sci 27:797–803

Smith C, Simpson SP (1986) Use of genetic polymorphisms in livestock improvement. J Anim Breed Genet 103:203–217

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

Tanksley SD, Medino-Filho H, Rick CM (1982) Use of naturally occurring enzyme variation to detect and map genes controlling quantitative traits in an inter-specific backcross of tomato. Heredity 49:11–25