Genet. Res., Camb. (1999), **73**, *pp.* 177–184. *With* 1 *figure.* Printed in the United Kingdom © 1999 Cambridge University Press

Allele frequency changes in artificial selection experiments: statistical power and precision of QTL mapping

YUSEOB KIM* AND WOLFGANG STEPHAN

Department of Biology, University of Rochester, Rochester, NY 14627-0211, USA

(*Received 10 June 1998 and in revised form 23 September and 22 October 1998*)

Summary

A simple mathematical model of genic directional selection is developed to study frequency changes of genetic marker alleles that are partially linked to a quantitative trait locus (QTL) under artificial selection. The effects of population size, number of generations of artificial selection, recombination between marker locus and QTL, and the strength of selection on the change in allele frequency are analysed by the diffusion equation approach and by stimulation. Using these results, we investigate the power of statistical tests for the detection of QTLs based on the observation of significant marker allele frequency changes in selection experiments. The probability of inferring the correct location of a QTL is also obtained.

1. Introduction

Artificial selection experiments usually reveal a considerable amount of genetic variation that gives rise to extreme phenotypes not usually found in natural populations (Falconer & Mackay, 1996). The response results from genotypic changes at quantitative trait loci (QTLs) under selection. In conventional artificial selection experiments, only the change in the population mean of a trait is observed. Thus, it is difficult fully to understand the genetic basis of the response to selection. Recently, an experimental approach was proposed to identify the number, locations and effects of QTLs responsible for phenotype changes observed in a population under artificial selection (Lebowitz *et al*., 1987; Keightley & Bulfield, 1993; Keightley *et al*., 1996). This approach followed earlier work by Garnett & Falconer (1975). It is based on the observation of allele frequency changes at genetic markers that are assumed to be closely linked to the loci under selection. The frequency of a selected allele will increase or decrease depending on the phenotypic value of the QTL and the direction of artificial selection. Therefore, a QTL of sufficiently large effect can be detected by frequency changes of marker alleles that significantly deviate from the respective frequencies in the base population. A systematic experimental approach based on this principle is now possible due to dense maps of molecular genetic markers available in many animals and plants.

For the experiments mentioned above, an appropriate statistical method should be developed to ensure that the observed frequency changes are due to selection rather than random genetic drift. Keightley *et al*. (1996) used a maximum likelihood method to detect the effect of QTLs affecting mouse body weight. Their method depends on extensive Monte Carlo simulations of a specific selection experiment. Ollivier *et al*. (1997) proposed a different statistical analysis to estimate the average effect of a marker allele that is completely associated with a QTL. In this paper, a mathematical model of allele frequency changes in an artificial selection experiment is developed to evaluate the statistical power for detection and mapping of QTLs. We compute the expected mean and variance of the distribution of marker allele frequency changes under the null hypothesis (genetic drift) and the alternative hypothesis (directional selection). We consider an artificial selection experiment in which the base population is the F2 population of a cross between two inbred lines having different phenotypic values. Only one QTL is assumed to be linked to a given marker. Selection is assumed to be unidirectional.

* Corresponding author. Telephone: +1 (716) 275 5013. Fax: +1 (716) 275 2070. e-mail: yuse@troi.cc.rochester.edu.

https://doi.org/10.1017/S0016672398003711 Published online by Cambridge University Press

2. Theory

(i) *Model*

In a cross between two inbred lines, one line is assumed to be homozygous for allele B at a selected locus and allele M at a marker locus. In contrast to the stimulation model of Keightley *et al.* (1996), we consider only a single marker linked to the QTL. The other line is homozygous for b and m at the corresponding loci. The two loci are linked; the recombination fraction is c. Artificial selection begins with the F2 population from the cross between the two inbred lines. In the F2 generation ($t = 0$), the expected frequencies (x_1, x_2, x_3 and x_4) of chromosomes BM, Bm, bM and bm are $(1-c)/2$, $c/2$, $c/2$ and $(1-c)/2$, respectively. Therefore, the expected frequency of allele B, p, and that of allele M, q, assume the value $p_0 = q_0 = 0{\cdot}5$ at $t = 0$. (In reality, the initial frequency at $t = 0$ may deviate from 0·5 due to sampling variance occurring from the F1 to the F2 generation. This will be taken into account later.) We assume that effective (diploid) population size N is constant throughout the selection experiment. For the QTL, a genic selection model is considered, in which individuals with genotype BB, Bb and bb have fitness $1+2s$, $1+s$ and 1, respectively.

(ii) *Diffusion approximation*

To analyse the effects of selection and drift on the trajectory of the marker allele, we use a diffusion equation method. To obtain an expression for the drift coefficient of the diffusion equation, we begin by considering the first moment of the frequency of the selected allele. Using equation (5.2.18) from Crow & Kimura (1970), the change in allele frequency from one generation to the next is

$$\Delta p_t = \frac{sp(1-p)}{1+2sp} \approx \frac{sp(1-p)}{1+2sp_0}.$$

This approximation is valid only as long as the frequency of B stays reasonably close to p_0. With $p_0 = 0{\cdot}5$, the solution in the continuous time approximation becomes

$$p_t = \frac{1}{1+\mathrm{e}^{-s't}}, \quad \text{with} \quad s' = \frac{s}{1+s}. \tag{1}$$

To obtain the change in the first moment of the marker allele frequency, we introduce the variables Q and R to denote the proportions of M in those chromosomes containing B and b, respectively. Then, x_1, x_2, x_3 and x_4 and x_4 are rewritten as pQ, $p(1-Q)$, $(1-p)R$ and $(1-p)(1-R)$, respectively (Maynard Smith & Haigh, 1974). Furthermore, $q_t = p_t Q_t + (1-p_t)R_t$ and $Q_0 = 1-c$ and $R_0 = c$. It follows from Maynard Smith & Haigh (1974) that

$$\frac{dQ_t}{dt} = c(1-p_t)(R_t - Q_t),$$

$$\frac{dR_t}{dt} = cp_t(Q_t - R_t),$$

and

$$Q_t - R_t = (Q_0 - R_0)\,\mathrm{e}^{-ct} = (1-2c)\,\mathrm{e}^{-ct}.$$

It follows from these equations that

$$\frac{dq_t}{dt} = (1-2c)\frac{dp_t}{dt}\mathrm{e}^{-ct}. \tag{2}$$

Using (1) and (2), we approximate the q_t process by a one-dimensional diffusion, such that the diffusion coefficient is given by $\sigma^2(q,t) = q(I-q)/2N$ and the drift coefficient $\mu(q,t)$ by the right-hand side of (2). Note that the drift coefficient is time-dependent. This takes into account that the underlying process is higher-dimensional because the marker locus is linked to the QTL under consideration. The assumptions made here ignore stochastic fluctuations in the variables which have been eliminated by this reduction procedure, i.e. the frequency of the selected allele and linkage disequilibrium. The validity of these assumptions is examined by simulation (see below). Second-order approximations that take stochastic fluctuations into account are difficult to treat mathematically (but for an example in which second-order approximations could be obtained in the diffusion equation of the reduced system see Stephan *et al.*, 1999).

(iii) *Moments of q_t*

Combining (1) and (2) leads to the first moment

$$q_t \approx 0{\cdot}5 + (1-2c)\left(\frac{\mathrm{e}^{-ct}}{1+\mathrm{e}^{-s't}} - 0{\cdot}5 + \int_0^t \frac{\mathrm{e}^{-c\tau}}{1+\mathrm{e}^{-s'\tau}}\,\mathrm{d}\tau\right). \tag{3}$$

This is the expected frequency of M at generation t given it was 0·5 at $t = 0$. The second moment of q_t is computed using the equation

$$\frac{d}{dt}E_t(f) = E_t\left(\mu(q,t)\frac{d}{dq}f + \tfrac{1}{2}\sigma^2(q,t)\frac{d^2}{dq^2}f\right), \tag{4}$$

where $f = q^2$ (Stephan *et al.*, 1992). To solve this differential equation we have to make the additional assumption that p_t changes in a linear fashion with time about $p_0 = 0{\cdot}5$ (instead of (1)). Thus,

$$\frac{dp_t}{dt} \approx \frac{s'}{4}.$$

This assumption is valid as long as st is in the order of 1. Then, $\mu(q, t)$ becomes independent of q or p, and depends only on t. Next we define

$$H(t) = \int_0^t (1-2c)\frac{dp_\tau}{d\tau}\mathrm{e}^{-c\tau}\mathrm{d}\tau.$$

Then,

$$E_t(q) = q_0 + H(t), \qquad (5)$$

and

$$\frac{dE_t(q^2)}{dt} = E_t\left\{\frac{dq}{dt}\frac{d(q^2)}{dq} + \frac{1}{2}\frac{q(1-q)}{2N}\frac{d^2(q^2)}{dq^2}\right\}$$
$$= 2\frac{dH(t)}{dt}E_t(q) + \frac{1}{2N}(E_t(q) - E_t(q^2))$$
$$= \left(2\frac{dH(t)}{dt} + \frac{1}{2N}\right)(q_0 + H(t)) - \frac{1}{2N}E_t(q^2).$$

The solution of this differential equation is

$$E_t(q^2) = \mathrm{e}^{-\frac{t}{2N}}E_0(q^2) + \frac{1-\mathrm{e}^{-\frac{t}{2N}}}{2}$$
$$+ H(t) + 2\mathrm{e}^{-\frac{t}{2N}}\int_0^t \mathrm{e}^{\frac{\tau}{2N}}\frac{dH(\tau)}{d\tau}H(\tau)\,d\tau.$$

Using (5) and $E_0(q^2) = q_0^2 = \frac{1}{4}$, we have

$$Var(q_t) = E_t(q^2) - \{E_t(q)\}^2$$
$$= \frac{1-\mathrm{e}^{-\frac{t}{2N}}}{4} + 2\mathrm{e}^{-\frac{t}{2N}}\int_0^t \mathrm{e}^{\frac{\tau}{2N}}\frac{dH(\tau)}{d\tau}H(\tau)\,d\tau - H(t)^2. \qquad (6)$$

To further evaluate this equation, we use the relationship

$$H(t) \approx \frac{s'(1-2c)}{4c}(1-\mathrm{e}^{-ct}).$$

If c is much larger than $1/2N$,

$$\int_0^t \mathrm{e}^{\frac{\tau}{2N}}\frac{dH(\tau)}{d\tau}H(\tau)\,d\tau \approx \frac{s'^2(1-2c)^2}{16c}\int_0^t (1-\mathrm{e}^{-c\tau})\,\mathrm{e}^{(\frac{1}{2N}-c)\,\tau}\,d\tau$$
$$\approx \frac{s'^2(1-2c)^2}{16c}\int_0^t (\mathrm{e}^{-c\tau} - \mathrm{e}^{-2c\tau})\,d\tau$$
$$= \frac{s'^2(1-2c)^2}{32c^2}(1-\mathrm{e}^{-ct})^2.$$

Therefore,

$$Var(q_t) \approx \frac{1-\mathrm{e}^{-\frac{t}{2N}}}{4}$$
$$- \frac{s'^2(1-2c)^2}{16}(1-\mathrm{e}^{-\frac{t}{2N}})\left(\frac{1-\mathrm{e}^{-ct}}{c}\right)^2. \qquad (7a)$$

If $s = 0$ or $c = 0{\cdot}5$, the right-hand side is reduced to $(1-\mathrm{e}^{-\frac{t}{2N}})/4$, which has also been obtained by other methods (Crow & Kimura, 1970, p. 328). To incorporate the sampling variance occurring between the F1 and F2 generations, we replace t in the right-hand side of (7*a*) with $t+1$. This is only an approximation because there is no deterministic change in the frequency of B between the F1 and F2 generations. Then,

$$Var(q_t) \approx \frac{1-\mathrm{e}^{-\frac{t+1}{2N}}}{4}$$
$$- \frac{s'^2(1-2c)^2}{16}(1-\mathrm{e}^{-\frac{t+1}{2N}})\left(\frac{1-\mathrm{e}^{-c(t+1)}}{c}\right)^2. \qquad (7b)$$

(iv) *Statistical test of marker frequency change*

If we assume that q_t is normally distributed, we can calculate the type I and type II errors of hypothesis testing:

$$\mathrm{H}_0 : s = 0 \text{ or } c = 0{\cdot}5, \quad \mathrm{H}_1 : s \neq 0 \text{ and } c \neq 0{\cdot}5.$$

Under H_0, q_t is normally distributed with mean, $\mu_0 = 0{\cdot}5$, and variance

$$\sigma_0{}^2 = \frac{1-\mathrm{e}^{-\frac{t+1}{2N}}}{4}.$$

Under H_1, the mean, μ_1, and variance, σ_1^2, are given by (3) and (7*b*). Thus, the power of rejecting H_0 is given by the probability

$$\mathrm{P}[|q_t - \mu_0| > z_{\alpha/2}\sigma_0]$$
$$= \mathrm{P}\left[Z > \frac{\mu_0 + z_{\alpha/2}\sigma_0 - \mu_1}{\sigma_1} \quad \text{or} \quad Z < \frac{\mu_0 - z_{\alpha/2}\sigma_0 - \mu_1}{\sigma_1}\right], \qquad (8)$$

where $Z \sim N(0, 1)$, $\alpha =$ probability of type I error.

(v) *Frequency difference of selected and marker loci*

When $s > 0$, p_t becomes greater than q_t on average. But due to random genetic drift, there is a certain probability that q_t will instead become greater than p_t. If one genetic marker is completely linked with a QTL and another one with recombination fraction c, the location of the QTL will be correctly inferred only when $p_t > q_t$. Therefore, the probability $\mathrm{P}[p_t > q_t]$ for a given c can be a useful measure of the precision of QTL mapping. Thus we investigate the distribution of $y = p_t - q_t = x_2 - x_3$ at generation t. The mean, $E(y) = E(p_t) - E(q_t)$, is given by (1) and (3). To obtain the variance of y, we again use (4). In this case, $f = y^2$, and y is assumed to follow a one-dimensional diffusion. Because simulation indicated that the variance of y does not change much by s as long as st is of the order

https://doi.org/10.1017/S0016672398003711 Published online by Cambridge University Press

of 1, we obtain the variance for the case $s = 0$. Then, as the drift parameter becomes zero,

$$\frac{d}{dt}E_t(y^2) = E_t\left(\tfrac{1}{2}\sigma^2(y,t)\frac{d^2}{dy^2}y^2\right),$$

where

$$\begin{aligned}\sigma^2(y,t) &= V^*(x_2 - x_3) = V^*(x_2) + V^*(x_3) \\ &\quad - 2Cov^*(x_2, x_3) \\ &= \frac{x_2(1-x_2)}{2N} + \frac{x_3(1-x_3)}{2N} - 2\left(-\frac{x_2 x_3}{2N}\right) \\ &= \frac{x_2 + x_3 - y^2}{2N}.\end{aligned}$$

V^* and Cov^* represent the sampling variance and covariance at each generation, given by the multinomial distribution. Then, (4) becomes

$$\frac{dE_t(y^2)}{dt} = \frac{1}{2N}E_t(x_2 + x_3) - \frac{1}{2N}E_t(y^2).$$

Therefore,

$$E_t(y^2) = e^{-\frac{t}{2N}}E_0(y^2) + \frac{e^{-\frac{t}{2N}}}{2N}\int_0^t e^{\frac{\tau}{2N}}E_\tau(x_2 + x_3)\,d\tau.$$

Using $E_0(y^2) = 0$ and $E_t(y) = 0$ for $s = 0$,

$$\begin{aligned}Var(y) &= E_t(y^2) - (E_t(y))^2 \\ &= \frac{e^{-\frac{t}{2N}}}{2N}\int_0^t e^{\frac{\tau}{2N}}E_\tau(x_2 + x_3)\,d\tau. \qquad (9)\end{aligned}$$

Then, using

$$\frac{dx_2}{dt} = \frac{dx_3}{dt} = cD_t,$$

where

$$\begin{aligned}D_t &= (x_1 x_4 - x_2 x_3)_t = D_0 e^{-ct}, \\ E_t(x_2 + x_3) &= E_0(x_2 + x_3) + 2D_0(1 - e^{-ct}) \\ &= c + \frac{1-2c}{2}(1 - e^{-ct}).\end{aligned}$$

Together with (9), this leads to

$$\begin{aligned}Var(y) &= \frac{e^{-\frac{t}{2N}}}{2N}\int_0^t e^{\frac{\tau}{2N}}\left\{c + \frac{1-2c}{2}(1 - e^{-c\tau})\right\}d\tau \\ &= \tfrac{1}{2}(1 - e^{-\frac{t}{2N}}) - \frac{1-2c}{2(1-2Nc)}(e^{(\frac{1}{2N}-c)\,t} - 1). \qquad (10a)\end{aligned}$$

Then, as with (7*b*), sampling drift between the F1 and F2 generations can be incorporated by replacing t with $t+1$. Therefore,

$$\begin{aligned}Var(y) = \sigma_y^2 &= \tfrac{1}{2}(1 - e^{-\frac{t+1}{2N}}) \\ &\quad - \frac{1-2c}{2(1-2Nc)}(e^{(\frac{1}{2N}-c)\,(t+1)} - 1). \qquad (10b)\end{aligned}$$

We also assume that y is approximately normally distributed. Then, the probability that p_t is larger than q_t is

$$\mathrm{P}[p_t > q_t] = \mathrm{P}[y > 0] \approx \mathrm{P}\left[Z > -\frac{E(y)}{\sigma_y}\right], \qquad (11)$$

where $Z \sim N(0, 1)$.

3. Simulation

Simulation of artificial selection experiments was performed to test the accuracy of our analytic approximations. In each generation, each of N individuals has two arrays (chromosomes) on which L loci are located. The first locus is the selected locus and the other $L-1$ loci are marker loci. At the beginning of the simulation, all N individuals are heterozygous at all loci (F1 population). Then, the next generation (F2, $t = 0$) is created by sampling N pairs of parents with replacement. One gamete is generated from each pair of F1 parents, allowing recombination to occur with probability c between adjacent loci. In subsequent generations the probability of being chosen as a parent is proportional to the fitness of the individual, namely, $1+2s$, $1+s$ and 1 for an individual having genotype *BB*, *Bb* and *bb*, respectively, at the first locus. This simulation procedure is identical to the fertility selection approximation of Keightley *et al.* (1996), except that we use $2s$ for the fitness difference between two homozygotes instead of s. In the tth generation the frequency of B at the selected locus and that of M at a marker locus are recorded. The distributions of p_t and q_t values were obtained by repeating this simulation 10000 times for each parameter set. The program was written in C and run on a PowerMac.

We evaluated the accuracy of our approximate formulas by comparing the numerical solution with the simulation results (Table 1). The approixmations are good as long as allele frequencies have not changed more than 0·25. But the solution for σ_y was rather inaccurate with small Nc ($N = 100$, $c = 0{\cdot}0196$). The weak dependence of σ_y on s is seen in the simulation.

Tables 2, 3 and Fig. 1 show the power of the statistical test where the presence of a QTL is confirmed by a significant change in the allele frequency of a linked marker. The distribution of q_t under the null hypothesis is assumed to be normal with mean 0·5 and variance σ_0^2. Then, we used a two-sided test with a 99% confidence level. Due to the restrictions of the mathematical models we have, the number of generations was adjusted so that st is less than 1·5. When s is large, the allele frequency and the power will increase rapidly and remain close to 1·0 after a relatively small number of generations. We are

https://doi.org/10.1017/S0016672398003711 Published online by Cambridge University Press

Table 1. *Comparison of analytic solutions and simulation results*

		$s = 0{\cdot}02$, $t = 20$		$s = 0{\cdot}06$, $t = 20$		$s = 0{\cdot}14$, $t = 9$		$s = 0{\cdot}22$, $t = 4$	
		Theory[a]	Simul.[b]	Theory	Simul.	Theory	Simul.	Theory	Simul.
$N = 100$	q_t	0·5770	0·5715	0·7056	0·6990	0·7222	0·7168	0·6599	0·6572
$d = 2$ cM	σ_1	0·1558	0·1540	0·1395	0·1382	0·0932	0·0935	0·0716	0·0727
	σ_y	0·0800	0·1016	0·0800	0·0968	0·0451	0·0526	0·0276	0·0310
$N = 550$	q_t	0·5770	0·5762	0·7056	0·7057	0·7222	0·7210	0·6599	0·6599
$d = 2$ cM	σ_1	0·0678	0·0675	0·0607	0·0597	0·0402	0·0398	0·0307	0·0310
	σ_y	0·0432	0·0436	0·0432	0·0418	0·0233	0·0227	0·0135	0·0131
$N = 250$	q_t	0·5433	0·5421	0·6168	0·6173	0·6588	0·6605	0·6280	0·6293
$d = 8$ cM	σ_1	0·1010	0·1003	0·0980	0·0974	0·0654	0·0658	0·0472	0·0473
	σ_y	0·1064	0·1066	0·1064	0·1008	0·0618	0·0586	0·0374	0·0359
$N = 250$	q_t	0·5968	0·5936	0·7561	0·7501	0·7512	0·7468	0·6729	0·6698
$d = 0$ cM[c]	σ_1	0·0992	0·0981	0·0816	0·0805	0·0555	0·0555	0·0445	0·0449

[a] The numerical solutions of (3), (7*b*) and (10*b*) for q_t, σ_1 and σ_y, respectively.
[b] The observed values of q_t, σ_1 and σ_y averaged over 10000 simulations for each parameter set.
[c] For σ_1, $c = 10^{-5}$ was used in (7*b*); σ_y is not applicable.

Table 2. *Power of QTL detection: effect of recombination* ($N = 250$, $\alpha = 0{\cdot}01$)

	$s = 0{\cdot}02$ $t = 20$		$s = 0{\cdot}06$ $t = 20$		$s = 0{\cdot}1$ $t = 15$		$s = 0{\cdot}18$ $t = 6$		$s = 0{\cdot}22$ $t = 4$	
d[a]	Theory[b]	Simul.[c]	Theory	Simul.	Theory	Simul.	Theory	Simul.	Theory	Simul.
0 cM	0·048	0·037	0·473	0·475	0·866	0·832	0·893	0·874	0·840	0·819
2 cM	0·033	0·026	0·266	0·266	0·613	0·608	0·793	0·783	0·754	0·743
4 cM	0·024	0·018	0·160	0·156	0·413	0·422	0·682	0·674	0·664	0·661
8 cM	0·017	0·013	0·071	0·065	0·197	0·204	0·474	0·490	0·494	0·506
16 cM	0·012	0·010	0·027	0·025	0·064	0·063	0·217	0·231	0·258	0·280

[a] Map distance (in centimorgans) between the QTL and the marker, using Haldane's map function.
[b] The power of rejecting H_0 ($s = 0$ or $c = 0{\cdot}5$), given by (8).
[c] The proportion of simulation results that yielded $q_t > 0{\cdot}5 + 2{\cdot}58\sigma_0$ or $q_t < 0{\cdot}5 - 2{\cdot}58\sigma_0$, out of 10000 replicates.

Table 3. *Power of QTL detection: effect of population size* ($d = 2$ *cM*, $\alpha = 0{\cdot}01$)

	$s = 0{\cdot}02$ $t = 20$		$s = 0{\cdot}06$ $t = 20$		$s = 0{\cdot}1$ $t = 15$		$s = 0{\cdot}14$ $t = 9$		$s = 0{\cdot}22$ $t = 4$	
N	Theory[a]	Simul.[b]	Theory	Simul.	Theory	Simul.	Theory	Simul.	Theory	Simul.
100	0·018	—	0·074	0·049	0·164	0·156	0·251	0·258	0·275	0·275
250	0·033	0·026	0·266	0·266	0·613	0·608	0·753	0·752	0·754	0·754
550	0·070	0·064	0·679	0·689	0·975	0·966	0·993	0·991	0·992	0·990
1000	0·138	0·141	0·949	0·948	1·000	1·000	1·000	1·000	1·000	1·000

[a] The power of rejecting H_0 ($s = 0$ or $c = 0{\cdot}5$), given by (8).
[b] The proportion of simulation results that yielded $q_t > 0{\cdot}5 + 2{\cdot}58\sigma_0$ or $q_t < 0{\cdot}5 - 2{\cdot}58\sigma_0$, out of 10000 replicates.

less interested in QTLs of this strong effect because they are expected to be detected by other simple experimental designs.

As expected, the power increases with increasing s, increasing N, decreasing c and increasing t. But with small s (0·02), improving the other parameter values does not substantially increase the power. The degree of power decline with increasing genetic distance depends on the number of generations (Table 2), since recombination between the selected and the marker locus increases with time. The power increases with t (Fig. 1) because q_t increases faster than σ_0. However, the rate of increase declines as the number of generations increases, because the allele frequency change of the selected locus slows down and the linkage disequilibrium between QTL and marker locus disappears.

We also checked whether the assumption of a normal distribution is valid for q_t. We measured the proportions of simulation runs corresponding to the

https://doi.org/10.1017/S0016672398003711 Published online by Cambridge University Press

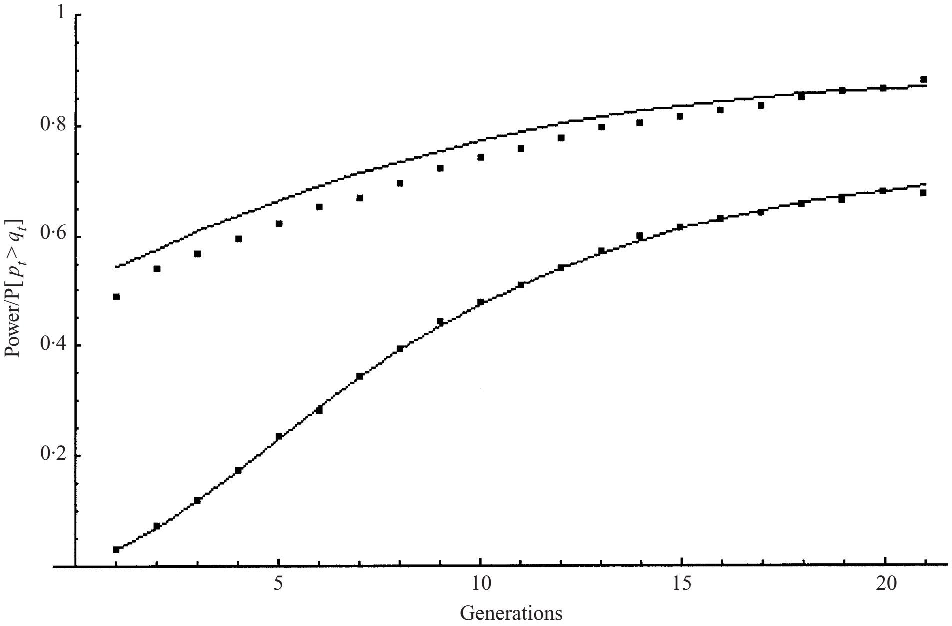

Fig. 1. The power and precision of QTL mapping with increasing number of generations, for $N = 250$, $s = 0{\cdot}1$, $c = 0{\cdot}0196$ (2 cM). Lower squares represent the power of test observed in the simulations (averaged over 10000 replicates). The analytic solution is shown by the lower line produced by (8). Upper squares and line represent the simulation and analytic result (equation (11)) of P[$p_t > q_t$], respectively.

Table 4. *Precision of QTL mapping* (P[$p_t > q_t$])

			$d = 2$ cM		$d = 4$ cM		$d = 8$ cM		$d = 16$ cM	
N	s	t	Theory[a]	Simul.[b]	Theory	Simul.	Theory	Simul.	Theory	Simul.
550	0·02	20	0·676	0·660	0·723	0·712	0·769	0·752	0·804	0·798
550	0·06	20	0·879	0·875	0·936	0·939	0·971	0·976	0·987	0·992
100	0·1	15	0·767	0·687	0·802	0·768	0·850	0·844	0·890	0·900
250	0·1	15	0·835	0·819	0·894	0·900	0·943	0·952	0·971	0·982
1000	0·1	15	0·965	0·973	0·992	0·996	0·999	1·000	1·000	1·000
250	0·22	4	0·748	0·700	0·816	0·786	0·886	0·887	0·941	0·930

[a] Probability given by (11).
[b] Proportion of simulation results that yielded $p_t > q_t$, out of 5000 replicates.

two tails of the empirical distribution of q_t (P[$|X - \overline{E(q_t)}| > z_{\alpha/2}\hat{\sigma}_1$], where X is the random variable following the empirical distribution, and $\overline{E(q_t)}$ and $\hat{\sigma}_1$ are the estimates of mean and standard deviation of the distribution, respectively. The result shows that the distribution has a slightly longer tail toward 0·5, and this skewness becomes greater as N becomes smaller (data not shown). However, the deviation from normality cannot be serious since the analytic solution for power agrees very well with the simulation results (Fig. 1).

We investigated the probability P[$p_t > q_t$] as a measure of the precision of QTL mapping (see Section 4). This is the probability that the location of a QTL is correctly inferred when one marker is completely linked to the QTL and another marker is linked with recombination fraction c. Table 4 shows that this probability increases with increasing c, increasing s and increasing N. P[$p_t > q_t$] also increases with increasing number of generations (Fig. 1), because of increased recombination between QTL and marker locus. The approximate solution for this probability, equation (11), generally overestimates P[$p_t > q_t$]. The empirical distribution of y obtained by simulation has a longer tail towards large values ($s > 0$), especially when the recombination fraction is small. This skewness causes the overestimation of our analytic solution.

4. Discussion

By assuming that changes in marker allele frequency follow a normal distribution and that only one selected locus is linked to the marker, we could apply the result of our diffusion approximation to the calculation of the power of detecting a QTL in an artificial selection

https://doi.org/10.1017/S0016672398003711 Published online by Cambridge University Press

experiment. Thus one can use our approximate formulas to determine the population size, the number of generations and the density of markers for the detection of a QTL of a desired effect. It is important to compare the efficiency of this approach with conventional QTL mapping methods, which test the marker–phenotype correlation in F2 or backcross populations (Lander & Botstein, 1989). Although the extensive evaluation of the power and efficiency of QTL mapping methods is beyond the scope of this paper, we can make a comparison with a simplified F2 design. For example, Soller *et al.* (1976) examined the power of experimental designs in crosses between inbred lines. They calculated that, with complete linkage between a marker and a QTL, 1050 F2 offspring are required to detect a QTL having a proportionate effect, $2a/\sigma = 0{\cdot}282$ ($2a =$ expected phenotypic difference between two homozygote classes, $\sigma =$ phenotypic standard deviation within a class), when type I error (α) $= 0{\cdot}05$ and type II error (β) $= 0{\cdot}1$. In order to make a comparison, we use an approximate formula for determining the selection coefficient in selection experiments, $2s = i(2a/\sigma)$ (Falconer & Mackay, 1996, p. 200), where i is the intensity of selection. Then, assuming $i = 1{\cdot}0$ (40% of population surviving at each generation: Falconer & Mackay, 1996, p. 190), we find the selection coefficient of the QTL is approximately 0·14. Our analytic formula indicates that, in this case, artificial selection with $N = 90$ and $t = 14$ or $N = 229$ and $t = 7$ is required to detect this QTL that is completely linked to the marker.

When there is recombination between a marker and a QTL, sample size increases by $1/(1-2c)^2$ in an F2 design. Therefore, if the marker is 5 cM from the QTL, the number of offspring required in the F2 design described above increases to 1282 (the use of interval mapping reduces this increase in sample size: Lander & Botstein, 1989). In the case of artificial selection with $t = 14$, however, N increases from 90 to 293 for detecting the same QTL. Therefore, although the number of offspring to be genotyped can be much smaller in QTL mapping by artificial selection, the decline in power due to recombination is more serious than in an F2 design. The advantage over F2 or backcross designs (in terms of power) is maximized when one uses a very dense map of markers.

The number of recombination events between a marker and a QTL is the most important factor in determining the resolution of any QTL mapping design. We can thus expect better resolution of QTL location in an artificial selection experiment than in an F2 or backcross design. In the experiment using multiple linked markers, the location of a genetic marker which exhibits the largest frequency change will provide information on the location of QTL in question, although this cannot be a clear criterion of QTL location. The probability that the highest deviation in frequency is observed in a false marker (which is not the closest to the QTL) will increase as the genetic distances between markers become smaller. We can express the precision of this QTL mapping as the maximum possible density of markers above which the probability of correctly inferring the closest marker to the QTL cannot exceed a certain value (the highest value is obtained when one of the markers happens to be completely linked to the QTL). Then, the probability $P[p_t > q_t]$ we obtained will provide the maximum possible density for a given experimental condition. For example, if we want a maximum accuracy of 90% of inferring the closest marker to the QTL of $s = 0{\cdot}1$, the average distance between markers should not be less than 4 cM when $N = 240$ and $t = 15$ (Table 4).

We conducted both analytic and simulation studies of artificial selection experiments in which effective population size (less than the number of surviving offsprings at each generation) is at least 100. Although most selection experiments are conducted with smaller populations (usually less than 50), there are difficulties in applying our result to a small population. First, the variance of q_t becomes large in small populations so that, after a certain number of generations, $p_0 + z_{\alpha/2}\sigma_0$ exceeds 1·0 and thus the statistical test is impossible. For example, with $t = 15$ and $\alpha = 0{\cdot}01$, N cannot be lower than 50. Secondly, the distribution of q_t becomes more skewed as N gets smaller. Besides these reasons, the power of detecting a QTL is not high enough to be practically important when N is below 100 (Table 3). Therefore, the application of our result will be limited to large-scale selection experiments such as those carried out in *Drosophila* (Weber, 1996).

We thank Bill Hill, Allen Orr and two reviewers for helpful suggestions on a previous version of this paper. This research was supported in part by NSF grant DEB-9407226/9896179 to W.S. and a Caspari Fellowship of the University of Rochester to Y.K.

References

Crow, J. & Kimura, M. (1970). *Introduction to Population Genetics Theory*. New York: Harper & Row.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edn. London: Longman.

Garnett, I. & Falconer, D. S. (1975). Protein variation in strains of mice differing in body size. *Genetical Research* **25**, 45–57.

Keightley, P. D. & Bulfield, G. (1993). Detection of quantitative trait loci from frequency changes of marker alleles under selection. *Genetical Research* **62**, 195–203.

Keightley, P. D., Hardge, T., May, L. & Bulfield, G. (1996). A genetic map of quantitative trait loci for body weight in the mouse. *Genetics* **142**, 227–235.

Lander, E. S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Lebowitz, R. J., Soller, M. & Beckman, J. S. (1987). Trait-

https://doi.org/10.1017/S0016672398003711 Published online by Cambridge University Press

based analysis for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics* **73**, 556–562.

Maynard Smith, J. & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research* **23**, 23–35.

Ollivier, L., Messer, L. A., Rothschild, M. F. & Legault, C. (1997). The use of selection experiments for detecting quantitative trait loci. *Genetical Research* **69**, 227–232.

Soller, M., Brody, T. & Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47**, 35–39.

Stephan, W., Wiehe, T. H. E. & Lenz, M. W. (1992). The effect of strongly selected substitutions on neutral polymorphism: analytical results based on in diffusion theory. *Theoretical Population Biology* **41**, 237–254.

Stephan, W., Charlesworth, B. & McVean, G. (1999). The effect of background selection at a single locus on weakly selected, partially linked variants. *Genetical Research* **73**, 133–146.

Weber, K. E. (1996). Large genetic change at small fitness cost in large population of *Drosophila melanogaster* selected for wind tunnel flight: rethinking fitness surfaces. *Genetics* **144**, 205–213.

https://doi.org/10.1017/S0016672398003711 Published online by Cambridge University Press