Genet. Res., Camb. (1998), **72**, *pp.* 55–58. Printed in the United Kingdom © 1998 Cambridge University Press

SHORT PAPER

A note on the theory of artificial selection in finite populations and application to QTL detection by bulk segregant analysis

WILLIAM G. HILL*

Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK

(*Received 19 January 1998 and in revised form 16 March 1998*)

Summary

Formulae are given for computing the distribution of numbers of selected individuals of each genotype and thus change in gene frequency at a locus with a large effect on a quantitative trait under truncation selection in a finite population. Results are illustrated with respect to use of selection for quantitative trait locus (QTL) detection, specifically by bulk segregant analysis with linked markers, for which probabilities that selected samples will comprise almost all one genotypic class are computed.

In an artificial selection programme a number of individuals (M) are recorded for a quantitative trait (which may be an index of several traits), and a number of these (N) are selected as parents of the next generation (M and N may, or may not, be the same in each sex). Selection can also be based on performance of relatives of the individuals, but here selection is assumed to be by truncation on individual phenotype.

The expected change in mean genotypic value of the trait can be computed using order statistics, which take account of finite numbers (e.g. Falconer & Mackay, 1996), and the change in gene frequency at any locus that affects the trait can be computed (at least approximately) similarly. As the numbers of parents are finite, however, the changes in mean of the trait and in gene frequency are also influenced by stochastic factors. The simplest procedure for computing the distribution of the change in gene frequency is to assume that selection induces a fitness differential at the locus and to use standard methods for stochastic processes for directional changes in gene frequency (e.g. Robertson, 1960). Precise methods can be adopted, however, which make use of order statistic methods to compute the probability distribution of the number of individuals of each genotype among the selected group (Hill, 1969).

During a recent analysis of the fate of genes in a recurrent backcrossing programme with selection (Hill, 1998), it became apparent that a simpler formulation than given previously (Hill, 1969) was possible that considerably reduces complexity and computation times, and illustrates more clearly the process of truncation selection as it affects an individual locus. In this note the formula is merely spelled out and applied to a topic in quantitative trait locus (QTL) mapping.

For simplicity, first consider a two-state model (e.g. haploid or backcross population) with two (geno)types A_1 and A_2, in which A_1 has a mean genotypic value a phenotypic standard deviations greater than A_2, and in which A_1, and A_2 are sampled with frequencies q_1 and $q_2 = 1-q_1$, respectively. Let $\Phi(x)$ denote the distribution function and $\phi(x)$ the density function of the standardized normal distribution. Among the M individuals available for selection, the number, m_1, that are A_1 has a binomial (M, q_1) distribution, i.e. $\binom{M}{m_1} q_1^{m_1} q_2^{m_2}$. The number of alternative ways of taking, among the highest scoring N, n_1 of type A_1 from m_1 and $n_2 = N-n_1$ of type A_2 from $m_2 = M-m_1$ is $\binom{m_1}{n_1}\binom{m_2}{n_2}$. Assume first that the lowest-scoring selected individual has phenotypic value x and is one of the $n_1 A_1$ individuals. Therefore, among those selected there are a further n_1-1 individuals of type A_1, each with probability $1-\Phi(x)$, and n_2 of type A_2, each with probability $1-\Phi(x+a)$, that have higher phenotype; and there are m_1-n_1 of type A_1 and m_2-n_2 of type A_2 that have lower phenotype and are

* Telephone: +44 (0)131 650 5705. Fax: +44 (0)131 650 6564. e-mail: w.g.hill@ed.ac.uk.

Table 1. *Exact probabilities $P(n_1, n_2)$ of selecting n_1 A_1 and n_2 A_2 alleles from N selected out of M recorded, for two classes (e.g. a backcross) for a gene A_1 with effect a standard deviations, expected frequency $q_1 = 1/2$ among those recorded and q_1' among those selected*

		$M = 50$	$N = 10$					$M = 100$	$N = 10$	
n_1	a:	2·0	1·5	1·0	0·5	0·25	0·0	2·0	1·0	0·5
10		0·7185	0·4033	0·1272	0·0180	0·0048	0·0010	0·8827	0·2378	0·0318
9		0·2345	0·3708	0·2840	0·0883	0·0336	0·0098	0·1100	0·3620	0·1300
7–8		0·0464	0·2145	0·4798	0·4543	0·3078	0·1611	0·0073	0·3621	0·5057
4–6		0·0005	0·0113	0·1081	0·4192	0·5869	0·6562	0·0000	0·0379	0·3230
0–3		0·0000	0·0000	0·0008	0·0202	0·0670	0·1719	0·0000	0·0001	0·0096
q_1'		0·9665	0·9105	0·8098	0·6672	0·5852	0·5000	0·9875	0·8645	0·7069
		$M = 100$	$N = 20$					$M = 50$	$N = 5$	
n_1	a:	2·0	1·0	0·5		n_1	a:	2·0	1·0	0·5
19–20		0·8717	0·0941	0·0037		5		0·9317	0·4754	0·1746
17–18		0·1240	0·3783	0·0606		4		0·0659	0·3764	0·3625
13–16		0·0043	0·5027	0·6041		2–3		0·0024	0·1462	0·4330
7–12		0·0000	0·0248	0·3307		0–1		0·0000	0·0020	0·0299
0–6		0·0000	0·0000	0·0009						
q_1'		0·9693	0·8129	0·6689		q_1'		0·9859	0·8603	0·7041

rejected. Summing over the values of m_1 and integrating over x, the probability $P(n_1, n_2; A_1)$ that n_1 individuals of type A_1 are selected with an A_1 the poorest of these is

$$P(n_1, n_2; A_1) = \sum_{m_1=n_1}^{M-n_2} \binom{M}{m_1} q_1^{m_1} q_2^{m_2} \binom{m_1}{n_1} \binom{m_2}{n_2} \times n_1 \int_{-\infty}^{\infty} [1-\Phi(x)]^{n_1-1} [1-\Phi(x+a)]^{n_2} \times [\Phi(x)]^{m_1-n_1} [\Phi(x+a)]^{m_2-n_2} \phi(x)\, dx,$$

which reduces to

$$P(n_1, n_2; A_1) = \frac{M!\, q_1^{n_1} q_2^{n_2}}{(M-N)!\, n_1!\, n_2!} \times n_1 \int_{-\infty}^{\infty} [1-\Phi(x)]^{n_1-1} [1-\Phi(x+a)]^{n_2} \times [q_1 \Phi(x) + q_2 \Phi(x+a)]^{M-N} \phi(x)\, dx.$$

In previous analyses, the two steps of sampling M individuals and selecting N of them were not combined, and the summation over m_1 within the integral not undertaken (Hill, 1969); or selection of only $N = 1$ individual was considered (Hill, 1998). The probability $P(n_1, n_2; A_2)$ that an A_2 is the poorest of the n selected individuals follows by substitution; and the overall probability that n_1 of type A_1 are selected is $P(n_1, n_2) = P(n_1, n_2; A_1) + P(n_1, n_2; A_2)$.

Hence consider the general case where there are k different genotypes, of which the jth has genotypic value a_j phenotypic standard deviations, relative to some approximate zero mean, and frequency q_j. The probability $P(n_1, \ldots, n_j, \ldots, n_k) = P(\mathbf{n})$ that, for $j = 1, \ldots, k$, there are n_j selected of genotype j, where $\Sigma n_j = N$, out of a total of M recorded is therefore

$$P(\mathbf{n}) = \frac{M!}{(M-N)!} \left(\prod_j \frac{q_j^{n_i}}{n_j!} \right) \int_{-\infty}^{\infty} \left\{ \prod_j [1-\Phi(x-a_j)]^{n_j} \right\} \times [\textstyle\sum_j q_j \Phi(x-a_j)]^{M-N} \left\{ \sum_j \frac{n_j \phi(x-a_j)}{1-\Phi(x-a_j)} \right\} dx. \quad (1)$$

It can be shown that, as must be the case, the probabilities of all possible outcomes sum to one, and that if $a_j = 0$ for all j, (1) reduces to the multinomial distribution, $P(\mathbf{n}) = N! \prod_j (q_j^{n_j}/n_j!)$.

These results can be used to compute the mean and distribution of change in gene frequency from truncation selection. Examples are given in Table 1 from results computed for a different purpose. With selection of $N = 10$ from $M = 50$, a two-genotype model with $q_1 = 0{\cdot}5$ and $a = 0{\cdot}5$, the expected gene frequency in selected individuals is $q_1' = 0{\cdot}6672$. An approximate prediction (Robertson, 1960) is $q_1' = q_1 + iaq_1 q_2 = 0{\cdot}6715$, where $i = 1{\cdot}372$ is the selection intensity from order statistic tables (Falconer & Mackay, 1996). Whilst the approximation is satisfactory in this example, for $a = 2$ it predicts $q_1' > 1$. Further analyses conducted previously showed that approximations based on relating gene effects to fitness differences could be used to analyse long-term changes in gene frequency from artificial selection in finite populations, except when gene effects are large

https://doi.org/10.1017/S0016672398003334 Published online by Cambridge University Press

and selection very intense (Hill, 1969). As no family effects are included in the model, the results apply exactly only when there is selection within families or there are no other loci affecting the trait and no family environmental effects.

QTL detection and bulk segregant analysis

Selection on a quantitative trait can be used to identify QTL from the change in gene frequency at putative QTL or at closely marker genes in linkage disequilibrium, for example following a cross (Lebowitz *et al*., 1987; Ollivier *et al*., 1997). The formulae given here can be used to predict the power and efficiency of such methods more precisely and to generalize recurrent backcrossing and selection methods (Hill, 1998).

A technique that has been proposed for efficiently identifying the location of a QTL is that of bulk segregant analysis (Michelmore *et al*., 1991), in which an F_1 cross of two inbred lines is backcrossed to one of the parent lines, and a group of high-scoring and a group of low-scoring individuals for a trait of interest are selected. DNA from members of each selected group is pooled, and typed for large numbers of markers. Provided a QTL has a sufficiently large effect that there is a high probability that the high and low pools each comprise almost all individuals of the same genotype, then the two should differ clearly in lanes on a gel for markers closely linked to the QTL. Thus the location requires only two samples (per replicate) for each marker. The results derived here can be used to show the probability that particular numbers, 0, 1,…, of individuals of each genotype are found in each selected pool, and thus the requirements for discriminating among bands on a gel of different intensity. Some examples are given in Table 1 (by integrating (1) using Simpson's rule), for pools of 50 or 100 individuals having equal expected frequencies of the two types ($q_1 = 1/2$) and selected pools of 5, 10 or 20 individuals. It is shown that unless the QTL, assumed to be unlinked to other QTL, has an effect of almost 2 SD, there will be considerable mixing.

The results in Table 1 apply to a marker locus located *at* the QTL, i.e. with no crossovers between marker and QTL and complete linkage disequilibrium between them, as would be the case if the parental lines are inbred. Assume that crossovers occur with probability r between the QTL (locus A) and the marker (locus B, with alleles B_1 and B_2 initially in coupling with alleles A_1 and A_2, respectively) and let $P^*(n_1, n_2)$ denote the probability there are n_1 and n_2

Table 2. *Exact probabilities $P^*(n_1, n_2)$ of obtaining n_1 B_1 and $N-n_2$ B_2 alleles and expected frequency (p_1') at a marker locus in a sample of N selected out of M recorded, for two classes (e.g. a backcross) for a marker locus initially in coupling and linked, with recombination fraction r, to a QTL with frequency 1/2 and effect a standard deviations*

		$M = 50$	$N = 10$	$a = 2$					
n_1	r:	0	0·005	0·01	0·02	0·05	0·1	0·2	0·5
10		0·7185	0·6845	0·6519	0·5910	0·4376	0·2598	0·0837	0·0010
9		0·2345	0·2579	0·2787	0·3135	0·3730	0·3726	0·2351	0·0098
7–8		0·0465	0·0569	0·0684	0·0938	0·1828	0·3379	0·5219	0·1611
4–6		0·0005	0·0007	0·0010	0·0017	0·0066	0·0296	0·1577	0·6562
0–3		0·0000	0·0000	0·0000	0·0000	0·0000	0·0001	0·0016	0·1719
p_1'		0·9665	0·9618	0·9572	0·9479	0·9199	0·8732	0·7799	0·5000
		$M = 50$	$N = 10$	$a = 0·5$		$M = 50$	$N = 10$	$a = 1$	
n_1	r:	0	0·02	0·05	0·1	0	0·02	0·05	0·1
10		0·0180	0·0163	0·0139	0·0107	0·1272	0·1087	0·0856	0·0567
9		0·0883	0·0822	0·0737	0·0610	0·2840	0·2641	0·2340	0·1858
7–8		0·4543	0·4440	0·4277	0·3986	0·4798	0·4974	0·5159	0·5262
4–6		0·4192	0·4350	0·4582	0·4955	0·1081	0·1285	0·1625	0·2271
0–3		0·0202	0·0226	0·0265	0·0342	0·0008	0·0012	0·0020	0·0041
p_1'		0·6672	0·6605	0·6504	0·6337	0·8098	0·7974	0·7788	0·7478
		$M = 50$	$N = 5$	$a = 1$		$M = 50$	$N = 5$	$a = 2$	
n_1	r:	0	0·02	0·05	0·1	0	0·02	0·05	0·1
5		0·4754	0·4367	0·3835	0·3064	0·9317	0·8434	0·7236	0·5545
4		0·3764	0·3893	0·4016	0·4063	0·0659	0·1457	0·2414	0·3468
2–3		0·1462	0·1712	0·2104	0·2785	0·0024	0·0109	0·0348	0·0980
0–1		0·0020	0·0028	0·0045	0·0088	0·0000	0·0000	0·0001	0·0007
p_1'		0·8603	0·8459	0·8243	0·7883	0·9858	0·9664	0·9373	0·8887

https://doi.org/10.1017/S0016672398003334 Published online by Cambridge University Press

individuals with marker alleles B_1 and B_2 in the sample. Then,

$$P^*(n_1, n_2) = (1-r)^N P(n_1, n_2) + r(1-r)^{N-1}[(n_1+1) \times P(n_1+1, n_2-1) + (n_2+1)\, P(n_1-1, n_2+1)]$$

+corresponding terms in 2, 3 or more crossovers.

(The items denote, respectively: no recombination, a recombination of one of the n_1+1 A_1B_1 haplotypes to A_1B_2, and a recombination of an A_2B_2 to A_2B_1). If the expected frequency of the QTL in the selected sample is q_1', that of the marker is $p_1' = (1-r)\, q_1' + r(1-q_1')$.

Examples are given in Table 2 of the distributions of the numbers of markers linked to QTL in bulk segregant samples. An approximation to the marker gene frequency in the selected group, assuming there is initially complete coupling between marker and QTL, is $p_1' = q_1 + iaq_1\, q_2(1-2r)$. For example, for $M = 50$, $N = 10$, $q_1 = p_1 = 1/2$ and $r = 0{\cdot}05$, the prediction is $p_1' = 0{\cdot}65435$. The binomial distribution can then be used to predict the number in each class: which for this example gives $P(10, 0) = 0{\cdot}65435^{10} = 0{\cdot}0144$, close to the exact value (0·0180, Table 2). This approximation is poor, however, if $ia > 1$.

The results in Table 2 show, as in Table 1, that bulk segregant analysis is likely to produce uniform selected groups for the QTL only if its effect approaches 2 SD, unless selection is very intense; as a rough guide $ia \geqslant 2$ is needed if the selected groups are to be unlikely to have few of the 'wrong' genotype. Discrimination using a marker is not greatly reduced, however, if it is less than about 5 cM from the QTL.

I am grateful to Peter Keightley for helpful comments.

References

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

Hill, W. G. (1969). On the theory of artificial selection in finite populations. *Genetical Research* **13**, 143–163.

Hill, W. G. (1998). Selection with recurrent backcrossing to develop congenic lines for QTL analysis. *Genetics* **148**, 1341–1452.

Lebowitz, R. J., Soller, M. & Beckmann, S. J. (1987). Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics* **73**, 556–562.

Michelmore, R. W., Paran, I. & Kessali, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genome region by using segregating populations. *Proceedings of the National Academy of Sciences of the USA* **88**, 9828–9832.

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.

Robertson, A. (1960). A theory of limits in artificial selection. *Proceedings of the Royal Society of London, Series B* **153**, 234–249.