Genet. Res., Camb. (2004), **83**, *pp.* 49–53. *With* 1 *figure.*
DOI: 10.1017/S0016672303006517 Printed in the United Kingdom

An improved formulation of marker heterozygosity in recurrent selection and backcross schemes

Z. W. LUO[1,2]* AND L. MA[2]
[1] *School of Biosciences, The University of Birmingham Edgbaston, Birmingham B15 2TT, UK*
[2] *Laboratory of Population & Quantitative Genetics, Institute of Genetics, Fudan University, Shanghai 200433, China*

(*Received 22 January 2003 and in revised form 6 August and 30 October 2003*)

Summary

This report presents a theoretical formulation for predicting heterozygosity of a putative marker locus linked to two quantitative trait loci (QTL) in a recurrent selection and backcross (RSB) scheme. Since the heterozygosity at any given marker locus maintained in such a breeding programme reflects its map location relative to QTL, the present study develops the theoretical analysis of the QTL mapping method that recently appeared in the literature. The formulae take into account selection, recombination and finite population size during the multiple-generation breeding scheme. The single-marker and two-QTL model was compared numerically with the model involving two linked marker loci and two QTL. Without recombination interference, the two models predict the same expected heterozygosity at the linked marker loci, indicating that the model is valid for predicting marker heterozygosity maintained at any loci in an RSB breeding scheme. The formulation is demonstrated numerically for several RSB schemes and its implications in developing a likelihood-based statistical framework for modeling the RSB experiments are discussed.

1. Introduction

Dissecting quantitative genetic variation into genes at a molecular level has been recognized as one of the greatest challenges facing geneticists in the twenty-first century. In the last two decades tremendous effort has been invested in mapping a wide spectrum of quantitative genetic variation in most important species, but the candidate regions inferred so far have been too coarse for accurate gene targeting. A method that is distinct from the current strategies of QTL mapping is the recurrent selection and backcross (RSB) method that was originally proposed by Wright (1952) and extended by Hill (1998). An RSB breeding scheme is initiated from crossing two inbred lines P_1 and P_2 that are assumed to be fixed for different alleles at marker loci and loci affecting a quantitative trait (QTL). A random sample of F_1 individuals are backcrossed to the recurrent parental line P_2 to produce n_F independent backcross families with a constant size of N. In each of these backcross families, n individuals are selected for phenotype of the non-recurrent parent P_1 to produce the next generation of the backcross families. The selection and backcrossing are repeated for T generations to create independent backcross families B_{iT} with $i=1, 2,\ldots, n_F$. In the RSB scheme, selection on a quantitative trait plays a role in maintaining the donor genome regions that contain the QTL, and the recipient genome is at the same time diluted by repeated backcrossing. Thus, mapping resolution of QTL with large effects may be greatly improved by appropriate implementation of the breeding schemes.

Instead of calculating the test statistic for locating QTL at any given chromosomal position, the RSB method surveys the level of heterozygosity at a given marker locus as an indication of its map location relative to QTL. Luo *et al.* (2002) presented a theoretical prediction of heterozygosity at any polymorphic marker locus on a chromosome bearing one or two genes affecting a quantitative trait (QTL) in an RSB breeding scheme. This demonstrated that numerical evaluation of the theoretical prediction under the three-loci (one marker and two QTL) model had to be restricted to circumstances where backcross family size was very small (e.g. not greater than 10). As the family size increases, calculation of the marker

* Corresponding author. Tel: +44 (0)121 4145404. Fax: +44 (0)121 4145925. e-mail: z.luo@bham.ac.uk

heterozygosity involves storing and computing a huge number of terms, resulting in extensive computer time and overloading of computer storage. The total number of terms to be calculated and stored in evaluating equation (7) in Luo *et al.* (2002) is mathematically equivalent to the number of different configurations of integers r_i $(i=1, 2,\ldots, K)$ such that $N=\sum_{i=1}^{K} r_i$, where N is the family size and $K=8$ in the present context, the number of all possible genotypes at the three loci in an RSB population. A general formula for this number is given in Feller (1957, p. 36) as

$$c(K, N)=\binom{N+K-1}{K-1}.$$

For example, $c(8, 50)=264\,385\,836$. In this short paper, we present a simplified but much more efficient method for calculating the marker heterozygosity.

2. Theoretical model and analyses

We first consider three linked loci: two affecting a quantitative trait and the other being a polymorphic marker that is devoid of any effect on the trait. In an RSB scheme described elsewhere (Hill, 1998; Luo *et al.*, 2002), there are at most two alleles segregating at each of these loci. The rationale of the model will be explained in the following discussion. Let the recipient and donor alleles be denoted by M and m at the marker locus, and by A and a and B and b at the first and second QTL respectively. There are eight possible genotypes in total at the three loci, and they are indexed by $i=1, 2,\ldots, 8$ corresponding to $Mm/Aa/Bb$, $Mm/Aa/bb$, $Mm/aa/Bb$, $Mm/aa/bb$, $mm/Aa/Bb$, $mm/Aa/bb$, $mm/aa/Bb$ and $mm/aa/bb$. Although there are three possible orders among the marker and QTL, their relative genetic distances may be defined by c_1 and c_2, recombination frequencies between loci 1 and 2 and between loci 2 and 3 respectively. Assuming there is no recombination interference, the recombination frequency between the first and the third loci will be $c=c_1(1-c_2)+(1-c_1)c_2$. For simplicity but without loss of generality, the design of an RSB scheme is characterized by four parameters: n_F, N, n and T.

Instead of working on the probability distribution of the number of individuals with various marker-QTL genotypes in an RSB family as in the previous studies (Hill, 1998; Luo *et al.*, 2002), we work out here the probability distribution of these genotypes within an RSB family at any number of generations. Let X_t (or X_t'), Y_t (or Y_t') and Z_t (or Z_t') be indicator variables respectively for the genotype at the marker locus, genotype at the two QTL and genotype at the marker and QTL at the generation t of an RSB family before (or after) selection. $X_t(X_t')=1$ or 2 for marker genotype Mm or mm, $Y_t(Y_t')=1, 2,\ldots, 4$ for four QTL genotypes Aa/Bb, Aa/bb, aa/Bb and aa/bb, and $Z_t(Z_t')=1, 2,\ldots, 8$ for the eight marker-QTL genotypes listed above. The probability distributions are defined as

$$\begin{aligned}
&f_{t,i}=Pr\{Z_t=i\}, \quad f'_{t,i}=Pr\{Z'_t=i\} \qquad i=1,2,\ldots,8\\
&p_{t,i}=Pr\{Y_t=i\}, \quad p'_{t,i}=Pr\{Y_t{}'=i\} \qquad i=1,2,3,4\\
&a_{t,ij}=Pr\{X_t=i|\,Y_t=j\} \quad i=1,2;\ j=1,2,3,4\\
&\theta_{ij}=Pr\{Z_t=i|Z'_{t-1}=j\} \quad 1\leqslant i,\ j\leqslant 8.
\end{aligned}$$

To model truncation selection of the quantitative trait in an RSB family with finite size N in which there are r_i individuals with the ith genotype at the linked QTL $(r_1+\cdots+r_4=N)$, we need to calculate the probability that among the n individuals selected from the N individuals there are s_i individuals with the ith genotype at the QTL $(s_1+\cdots+s_4=n)$. A general theory of phenotypic selection in finite populations was developed in Hill (1969) and extended to calculate the probability distribution of s_i in the RSB setting in Luo *et al.* (2002). In the present notation, the probability has a form

$$\begin{aligned}
\xi_{SR}&=Pr\{S=(s_1,s_2,s_3,s_4)|R=(r_1,r_2,r_3,r_4)\}\\
&=\prod_{k=1}^{4}\binom{r_k}{s_k}\sum_{i=1}^{4}s_i\int_{-\infty}^{\infty}[1-\Phi_i(x)]^{s_i-1}\prod_{j=1}^{4}\Phi_j(x)^{r_j-s_j}\prod_{j\neq i}^{4}[1-\Phi_j(x)]^{s_j}\phi_i(x)\mathrm{d}x,
\end{aligned}\qquad (1)$$

where $\phi_i(x)$ and $\Phi_i(x)$ are respectively the probability density function and the probability distribution function of a normal distribution with mean μ_i, which is the genetic mean of the ith QTL genotype, and variance 1.0.

To calculate the aforementioned probability distributions, we need to work out the transition probability of a given marker-QTL genotype from parental to offspring generation. The form of the transition probabilities, $\mathbf{\Theta}_X=(\theta_{ij})$, depends on the location of the marker locus relative to the linked QTL ($X=L$, M, R corresponding to the left, middle or right location of the marker locus). This can be done easily by three-loci segregation analysis in a backcross population. For example, $\mathbf{\Theta}_M$ has a form of

https://doi.org/10.1017/S0016672303006517 Published online by Cambridge University Press

$$\boldsymbol{\Theta}_M=\begin{pmatrix} (1-c_1)(1-c_2)/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ (1-c_1)c_2/2 & (1-c_1)/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ c_1(1-c_2)/2 & 0 & (1-c_1)/2 & 0 & 0 & 0 & 0 & 0 \\ c_1(1-c_2)/2 & c_1/2 & c_2/2 & 0.5 & 0 & 0 & 0 & 0 \\ c_1c_2/2 & 0 & 0 & 0 & [(1-c_1)(1-c_2)+c_1c_2]/2 & 0 & 0 & 0 \\ c_1c_2/2 & c_1/2 & 0 & 0 & [c_1(1-c_2)+(1-c_1)c_2]/2 & 0.5 & 0 & 0 \\ (1-c_1)c_2/2 & 0 & c_2/2 & 0 & [c_1(1-c_2)+(1-c_1)c_2]/2 & 0 & 0.5 & 0 \\ (1-c_1)(1-c_2)/2 & (1-c_1)/2 & (1-c_2)/2 & 0.5 & [(1-c_1)(1-c_2)+c_1c_2]/2 & 0.5 & 0.5 & 1 \end{pmatrix}.$$

With these probabilities and the initial probability distribution of the marker-QTL genotypes $f_{0,i}=Pr\{Z_0=i\}$, which are simply equal to the first column of $\boldsymbol{\Theta}_X$, the probability distribution, $f_{t,i}$, can be worked out for any t in the following recursive way:

$$a_{t,ij}=Pr\{X_t=i|Y_t=j\}=f_{t,4(i-1)+j}/[f_{t,j}+f_{t,4+j}], \tag{2}$$

$$p_{t,i}=Pr\{Y_t=i\}=f_{t,i}+f_{t,4+i}, \tag{3}$$

$$\begin{aligned} p'_{t,i}=Pr\{Y'_t=i\} &= \sum_{s_i=1}^{n} s_i Pr\{S_t=(s_1,s_2,s_3,s_4)\}/n \\ &= \frac{1}{n}\sum_{s_i=1}^{n} s_i\left[\sum_{0\leqslant r_i\leqslant N} Pr\{R_t=(r_1,r_2,r_3,r_4)\}Pr\{S_t=(s_1,s_2,s_3,s_4)|R_t=(r_1,r_2,r_3,r_4)\}\right] \\ &= \frac{1}{n}\sum_{s_i=1}^{n} s_i\left[\sum_{0\leqslant r_i\leqslant N} \frac{N!}{\prod_{k=1}^{4} r_k!}\prod_{i=1}^{4} p_{t,i}^{r_i}\times\xi_{SR}\right], \end{aligned} \tag{4}$$

$$f'_{t,i}=Pr\{Z'_t=i\}=Pr\{Y'_t=j\}Pr\{X'_t=i|Y'_t=j\}=p'_{t,j}a_{t,ij}, \tag{5}$$

$$f_{t+1,i}=Pr\{Z_{t+1}=i\}=\sum_{j=1}^{8} Pr\{Z'_t=j\}Pr\{Z_{t+1}=i|Z'_t=j\}=\sum_{j=1}^{8} f'_{t,j}\theta_{ij}. \tag{6}$$

Derivation of (4) implies the assumption that (r_1, r_2, r_3, r_4), the numbers of individuals with four different genotypes at the QTL within an RSB family, follows a multinomial distribution with parameters $q_{t,i}$ ($i=1$, 2, 3, 4) and N. The summation is over all possible $0\leqslant r_i\leqslant N$ ($i=1$, 2, 3, 4) such that $\sum_{i=1}^{4} r_i=N$, and this involves $c(4, N)=(N+1)(N+2)(N+3)/6$ terms. Because selection is only on the QTL other than the marker locus, the conditional probability of the marker genotype given a genotype at the QTL before selection will remain unchanged after selection. This entails derivation of (5). From the joint distribution of genotypes at the marker and QTL, the marker heterozygosity expected for the RSB scheme defined above is thus given by

$$h_t=f_{t,1}+f_{t,2}+f_{t,3}+f_{t,4}. \tag{7}$$

In the above model, the marker locus virtually plays a role of specifying a chromosomal position at which its relatedness in map location to the linked QTL is tested. By searching over all possible map positions of a chromosome, we may expect that heterozygosity reaches its peak values at the locations of the QTL. This thus suggests that heterozygosity at the multiple linked marker loci would be predicted separately from the present single marker locus model if recombination interference is assumed to be absent. To demonstrate this, we considered the situation where there were two markers linked to two QTL. Under the four-loci model, there are 16 possible genotypes segregating in an RSB breeding scheme. In addition, we need to take six possible configurations of the relative locations between the marker loci and QTL into account in calculating the one-generation genotypic transition probability $\boldsymbol{\Theta}_X=(\theta_{ij})$. The dynamics of frequencies of the 16 genotypes and thus heterozygosity at the two marker loci can be calculated by following the same principle as that of the above three-loci model. The detailed formulation of the four-loci model analysis is omitted here.

3. Numerical calculation

We demonstrated the theoretical analyses by numerically calculating the expected marker heterozygosity

https://doi.org/10.1017/S0016672303006517 Published online by Cambridge University Press

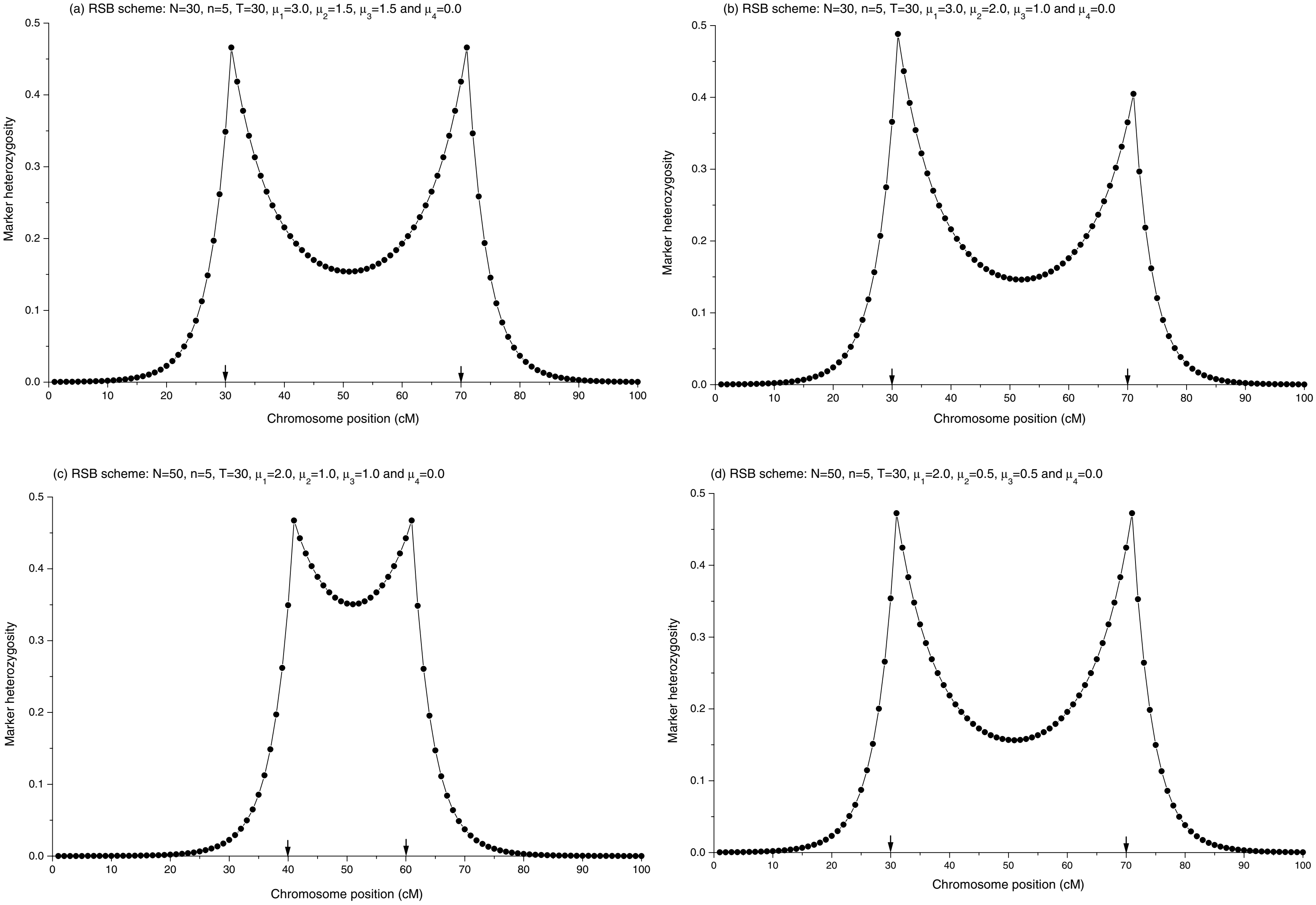

Fig. 1. Distribution of the expected heterozygosity over a chromosome of 100 cM which carries two linked QTL in four Recurrent Selection and Backcross (RSB) breeding schemes: (*a*) $N=30$, $n=5$, $T=30$, $\mu_1=3{\cdot}0$, $\mu_2=1{\cdot}5$, $\mu_3=1{\cdot}5$ and $\mu_4=0{\cdot}0$; (*b*) $N=30$, $n=5$, $T=30$, $\mu_1=3{\cdot}0$, $\mu_2=2{\cdot}0$, $\mu_3=1{\cdot}0$ and $\mu_4=0{\cdot}0$; (*c*) $N=50$, $n=5$, $T=30$, $\mu_1=2{\cdot}0$, $\mu_2=1{\cdot}0$, $\mu_3=1{\cdot}0$ and $\mu_4=0{\cdot}0$; and (*d*) $N=50$, $n=5$, $T=30$, $\mu_1=2{\cdot}0$, $\mu_2=0{\cdot}5$, $\mu_3=0{\cdot}5$ and $\mu_4=0{\cdot}0$. N is the family size, n is the number of individuals selected from each family, T is the number of consecutive selection and backcrossing of the breeding schemes, and μ_1, μ_2, μ_3 and μ_4 are means of the four genotypes at the linked QTL. Arrows indicate locations of the QTL.

from four RSB schemes defined by four sets of the parameters. Fig. 1 illustrates the distribution of marker heterozygosity predicted from these RSB breeding schemes. The marker heterozygosity was calculated at every 1 centimorgan (cM) over a chromosome of 100 cM that carried two linked QTL. It can be seen from the figure that consecutive selection and backcrossing in the RSB schemes have been effective in differentiating the recipient genome in the vicinity of the QTL from the donor genome backgrounds. Change in the level of heterozygosity at marker loci during the RSB breeding programme provides useful information on the QTL locations. In addition, the design parameters of RSB breeding schemes affect the sustainment of marker heterozygosity at the vicinity of QTL and the power in resolving the QTL, indicating that the RSB breeding schemes may be optimized to achieve an optimal resolving power of the QTL mapping.

We surveyed several arbitrarily chosen pairs of marker loci in Fig. 1*a* for their heterozygosity levels that were calculated from either the three-loci model or the four-loci model mentioned above, and found no difference between the two methods. Thus, the three-loci model may be used to assess genetic heterozygosity maintained at any genome location in an RSB scheme when recombination interference is absent.

The two-QTL model under the present study enables investigation of the power of the RSB method in resolving linked QTL and the effect of complex epistasis on QTL mapping. Following the same principle, the present analysis can readily be extended to multiple QTL but it will involve a substantial increase in computational demand. We considered a backcross family size of 50 in the above numerical demonstration of the theoretical analysis. This could be a realistic family size for most important model organisms, such as *Arabidopsis*, tomato, *Drosophila* or yeast, for which practising an RSB breeding scheme is feasible. Our experience shows that a modern personal computer is powerful enough to carry out the numerical computation in less than 10 hours of CPU-time even when the family size is doubled or tripled.

The present study does not merely serve as a solution to the theoretical gap of the previous research in the literature. The theoretical model discussed here can easily be modified into calculation of the conditional probability distribution of genotypes at any test chromosomal position given its flanking marker genotype in the RSB populations, entailing a likelihood framework for analysis of the RSB data as has been done in the conventional interval mapping technique (Lander & Botstein, 1989).

We thank Professor W. G. Hill for guiding our attention to Feller (1957) and his advice in revising the manuscript. The comments and criticisms from Professor M. J. Kearsey and anonymous reviewers have been helpful in improving the presentation of the paper and in removing several ambiguities. The present research was supported in part by a research grant from the UK Biotechnology and Biological Sciences Research Council. We are also grateful for financial support from China's '973', '863' programs and the National Natural Science Foundation.

References

Feller, W. (1957). *An Introduction to Probability Theory and its Applications*, vol. 1, 2nd edn. New York: Wiley.

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 quantitative trait loci analysis. *Genetics* **149**, 1341–1352.

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

Luo, Z. W., Wu, C. I. & Kearsey, M. J. (2002). Precision and high-resolution mapping of quantitative trait loci by use of recurrent selection, backcross or intercross schemes. *Genetics* **161**, 915–929.

Wright, S. (1952). The genetics of quantitative variability. In *Quantitative Inheritance* (eds. E. C. Reeve and C. H. Waddington), pp. 5–14. London: Her Majesty's Stationery Office.

https://doi.org/10.1017/S0016672303006517 Published online by Cambridge University Press