Genet. Res., Camb. (1994), **63**, *pp.* 39–47 *With 4 text-figures* Copyright © 1994 Cambridge University Press

Simulation of marker assisted selection in hybrid populations

A. GIMELFARB AND R. LANDE*
Department of Biology, University of Oregon, Eugene, Oregon 97403, USA

(*Received 22 June 1993 and in revised form 20 September 1993*)

Summary

A computer model is developed that simulates Marker Assisted Selection (MAS) in a population produced by a cross between two inbred lines. Selection is based on an index that incorporates both phenotypic and molecular information. Molecular markers contributing to the index and their relative weights are determined by multiple regression of individual phenotype on the markers. The model is applied to investigate the efficiency of MAS as affected by several factors including total number of markers in the genome, number of markers contributing to the index, population size and heritability of the character. It is demonstrated that selection based on genetic markers can effectively utilize the linkage disequilibrium between genetic markers and QTLs created by crossing inbred lines. Selection is more efficient if markers contributing to the index are re-evaluated each generation than if they are evaluated only once. Increasing the total number of markers in the genome as well as the number of markers contributing to the index does not necessarily result in a higher efficiency of selection. Moreover, too many markers may result in a weaker response to selection. Population size is shown to be the most important factor affecting the efficiency of MAS.

1. Introduction

Any method of selection that makes use of genetic markers requires finding markers that are associated with quantitative trait loci (QTLs) as well as estimating the contribution to the genotypic value of the trait by the QTLs associated with a marker (marker effect). Numerous works have been published recently on the genetic mapping of QTLs (e.g. Lander & Botstein, 1988; Paterson *et al.* 1988; Paterson *et al.* 1990). One of the stated goals of QTL mapping is to identify genetic markers that are closely linked to QTLs and, hence, can be used in selection for the trait. As stated by Zhang & Smith (1992), 'Eventually with very close linkage each QTL allele can be uniquely identified in selection, and selection will then be equivalent to selection on the QTLs themselves.'

Lande & Thompson (1990; also Lande, 1992) proposed a method of Marker Assisted Selection (MAS) which rather than actually mapping QTLs employs multiple regression of the phenotype on markers to identify a set of markers associated with QTLs as well as to estimate the marker effects. They recommend crossing two inbred lines to create linkage disequilibrium between genetic markers and QTLs that can be utilized by selection. They also suggest that a large number of markers should be included in the multiple regression in the generation immediately following the hybridization cross, but only those markers that yield the 'largest apparent additive effects' should be used in selection in this and in the subsequent generations. The main conclusion from the deterministic analysis in Lande & Thompson (1990) paper is that MAS based on an index incorporating marker effects together with phenotype can yield a greater response than selection based strictly on phenotype, provided there are sufficiently many markers and the population size is very large.

Zhang & Smith (1992) conducted computer simulations of selection in a population of 500 individuals of each sex. Three modes of selection were considered: based exclusively on the BLUP estimate (Kennedy & Sorensen, 1988) of the individual's genotypic value, based on an index incorporating the BLUP estimate as well as effects of genetic markers, and based exclusively on the marker effects. The highest response was by selection based on the combined index and the lowest by selection based exclusively on genetic markers.

In this paper we report results of computer

* Corresponding author.

https://doi.org/10.1017/S0016672300032067 Published online by Cambridge University Press

simulations aimed at investigating the effect of different factors (e.g. the number of available genetic markers, population size, heritability, etc.) on the effectiveness of MAS proposed by Lande & Thompson (1990) as compared to conventional mass selection based on phenotype.

2. Methods

The majority of simulations were conducted using the genetic map shown in Fig. 1. There were 25 diallelic quantitative trait loci (QTL) randomly distributed among 10 chromosomes of 100 cM each. Besides QTLs, each chromosome also had marker loci evenly spaced along the chromosome. The number of markers was the same (11 in Fig. 1) for all chromosomes, and a marker was always located at each end of a chromosome. The effects of the QTL alleles (shown in Fig. 1 under the corresponding loci) constitute the 'geometric series of variance contributions' as described by Lande & Thompson (1990). The actual 25 QTLs correspond to 10 'effective loci' (Lande, 1981). Simulations were also conducted using alternative maps: a map with 25 QTLs of the same effects as in Fig. 1 but distributed differently among the chromosomes, a map with 13 geometric series QTLs corresponding to 5 effective loci, and a map with 10 QTLs of equal effects. Some of the chromosomes in the last two maps had only marker loci on them, but no QTLs.

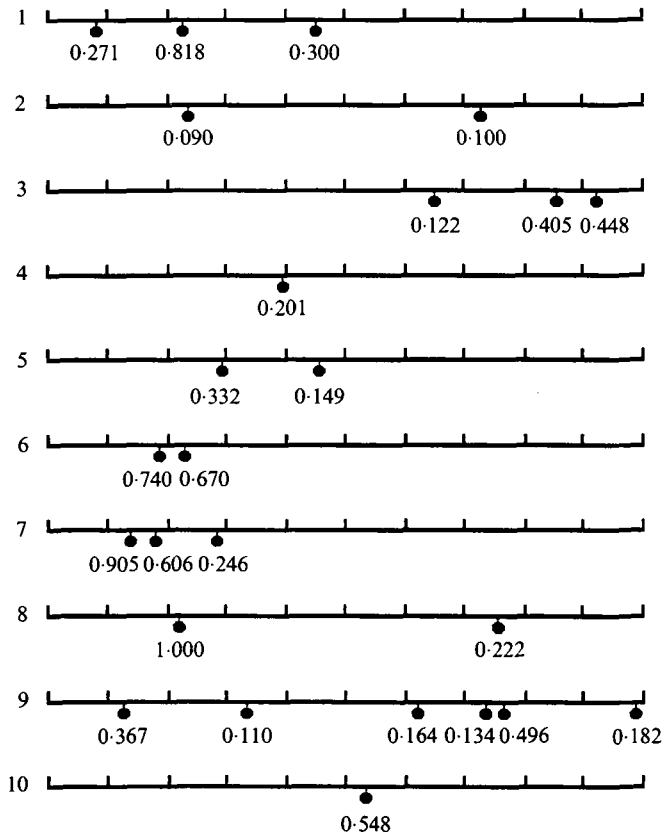

Fig. 1. Genetic map with 25 'geometric series' QTLs randomly distributed among 10 chromosomes and with 11 evenly spaced genetic markers per chromosome (numbers under QTLs indicate their additive effects on the character).

Recombination was simulated in the following manner. No more than two crossovers were allowed to occur between a pair of chromosomes of any length. The physical length of a chromosome was assumed to be n units such that one but no more than one chiasma can form within a unit. If the length of the chromosome in morgans is L, the distribution $C(i)$, of the probabilities for i chiasmata ($i = 0, 1, 2$) to be formed by a pair of such chromosomes can be written as

$$C(0) = (1-d)^n, \tag{1}$$

$$C(1) = nd(1-d)^{n-1}, \tag{2}$$

$$C(2) = 1 - C(0) - C(1), \tag{3}$$

where $d = 2L/n$ (2 accounts for the two pairs of chromatids). Consequently, the distribution $r(i)$ of the probabilities for i crossovers between the chromosomes can be obtained as

$$r(0) = C(0) + \tfrac{1}{2}C(1) + \tfrac{1}{4}C(2), \tag{4}$$

$$r(1) = \tfrac{1}{2}C(1) + \tfrac{1}{2}C(2), \tag{5}$$

$$r(2) = \tfrac{1}{4}C(2). \tag{6}$$

Given a crossover has occurred, its position along the chromosome was assumed random. The mapping function corresponding to this recombination process is practically indistinguishable from Haldane's mapping function (provided n is sufficiently large).

The initial population for each computer run was generated to represent an F1 cross between two inbred lines, i.e. all individuals were genetically identical and heterozygous for all loci. Two main types of F1 crosses with respect to the gametic phase of the QTLs were generated: 'total coupling' and 'total repulsion'. In total coupling phase, the effects of all alleles in the QTLs on a chromosome were in the same direction, e.g. $(+\,+\,+)$ on one chromosome and $(-\,-\,-)$ on its homologue. In total repulsion phase, the effects of adjacent QTL alleles on a chromosome were in opposite directions, e.g. $(+\,-\,+)$ and $(-\,+\,-)$. The marker loci in F1 crosses were always in total coupling phase for all runs.

The genotypic value, g, of an individual was assumed to be the sum of the allelic effects by all of the QTLs in the diploid genotype:

$$g = \sum_{i}^{n} (A_i' + A_i''), \tag{7}$$

where A_i' and A_i'' are the purely additive effects of maternal and paternal alleles at the ith QTL. The individual's phenotype, Z, was obtained as

$$Z = g + e, \tag{8}$$

where e is the environmental component normally distributed with zero mean and variance v_e. The value of the environmental variance was computed at the

https://doi.org/10.1017/S0016672300032067 Published online by Cambridge University Press

beginning of each run such as to yield a desired value of the heritability, $h^2 = v_g/(v_g+v_e)$, in generation F2, i.e.

$$v_e = v_g(1-h^2)/h^2, \quad (9)$$

where v_g is the expected genotypic variance in F2 which was estimated by the genotypic variance among 10000 offspring generated from the initial population. The value of the environmental variance computed at the beginning of a run remained unchanged for the total duration of the run (20 generations). The allelic effects of QTLs were scaled so as to yield the maximum and minimum of the genotypic value (by the two extreme homozygotes) equal to 10 and -10 units, respectively. Hence, the mean phenotypic value attainable under any selection could not exceed 10.

The offspring population in each generation was produced in the following manner. A male and a female chosen randomly from the parental pool of selected individuals were mated to produce one offspring. After that, they were returned to the parental pool. This was repeated until a specified total number of offspring had been obtained. Thus, the family size was variable.

Selection was based on the index proposed by Lande & Thompson (1990). A fixed proportion of individuals with the highest value of the index was selected for reproduction in each generation. The index of an individual with phenotype Z and molecular score M is

$$I = b_Z Z + b_M M. \quad (10)$$

Since only relative values of the coefficients in the index are relevant, the phenotypic coefficient, b_Z, was set to 1 in all our simulations, i.e. the index was computed as

$$I = Z + b_M M. \quad (11)$$

The individual's molecular score, M, is defined as

$$M = \sum_j c_j m_j, \quad (12)$$

where m_j is the number of alleles (0, 1 or 2) in the jth marker locus of the individual, and c_j is the additive effect associated with the marker, i.e. the coefficient of the multiple linear regression of the phenotype on the number of alleles at the marker locus. The summation is over all markers in the regression.

The computer program used in this study to simulate MAS was written to allow an investigation of different quantitative traits, including sex-dependent traits. Consequently, the regression of individual phenotype on markers was computed separately for the two sexes, even though traits considered in this report were sex-independent (with no sexual dimorphism). Therefore, individual males and females might have different molecular scores even if they were genetically identical.

Molecular scores of individuals in the first generation of any run were computed based on the regression utilizing all markers from all chromosomes in the genome. If the number of markers in the genome exceeds the number of individuals in the sample it is not possible to include all of the markers in a single regression. We therefore employed a two-stage procedure. In the first stage, a separate regression was fitted for each chromosome. The 'forward selection' procedure (Draper & Smith 1981) was employed to select from all of the markers on a chromosome only those five that made the highest contribution to the R^2 value of the regression. In the second stage, all previously selected 50 markers were thrown together and a regression was fitted utilizing all of them. The forward selection procedure was employed again to select a fixed number of markers with the highest contribution to the R^2 value. Such two-stage regressions were fitted in the first generation of all computer runs. As for the subsequent generations, in some runs the two-stage regression utilizing all markers in the genome was fitted in each generation, i.e. markers contributing to the molecular score differed between generations. Besides such runs 'with marker re-evaluation', runs were also conducted 'without marker re-evaluation'. Regressions in the subsequent generations of the latter runs utilized not all markers in the genome but only those selected in the first generation. A regression on all of these markers was fitted in one stage. Hence, markers contributing to the molecular score of a run without marker re-evaluation remained the same in each generation. The regression coefficients (additive effects) of the markers did change, however, since a new regression was fitted in each generation.

The molecular score coefficient, b_M, in the index (11) was computed as

$$b_M = (1/h^2-1)/(1-p_M), \quad (13)$$

where h^2 is the heritability and p_M is the proportion of the genetic variance accounted for by the markers in the regression. The latter can be expressed as

$$p_M = P_M/h^2, \quad (14)$$

where P_M is the proportion of the total phenotypic variance accounted for by the markers.

Since markers in the two-stage regression are not selected randomly, the standard squared correlation of the regression, R^2, overestimates P_M. The following method was employed to correct for the bias. Populations of individuals with phenotypes controlled exclusively by the environment (i.e. only with the markers but without QTLs in the genome) were generated for a given configuration of markers in the population. The two-stage regression of the phenotype on markers was fitted for each population. The average of the R^2 value over 40 such populations was used as the estimate of the bias in the R^2 for a regression with the same set of parameters but with the phenotype controlled by QTLs as well as environment. It turned out, however, that the correction

https://doi.org/10.1017/S0016672300032067 Published online by Cambridge University Press

Table 1. *Parameters used in simulations*

	Base	Alternative				
Markers on chromosome	11	3	6	21	51	101
Markers in selection index	6	3	9	12	15	20
Individuals of each sex	500	100	200	1000	3000	
Initial heritability	0·1	0·2	0·4			
Selection strength	25%	10%				

had practically no effect on the outcome of the simulations for population sizes greater than 100 individuals of each sex.

Parameters used in simulations are shown in Table 1. Most runs were conducted with parameter sets differing from the BASE set by only one alternative parameter at a time. However, some sets with two or more alternative parameters substituted in the BASE set were also investigated.

3. Results and discussion

The majority of the results reported here concern the efficiency of Marker Assisted Selection. Following Lande & Thompson (1990), the efficiency was calculated as a ratio of the response in the mean phenotype under MAS to the response under conventional phenotypic selection with the same set of parameters. A result for a given parameter set represents an average over replicated runs. The number of replicated runs depends on the population size: 40 runs for 100 and 200 individuals of each sex, 30 runs for 500 individuals of each sex and 20 runs for 1000 and 3000 individuals of each sex.

Runs with a given parameter set were conducted starting from initial populations in total coupling as well as in total repulsion gametic phase. Results for a cross between two real inbred lines should fall between these extremes, and this was confirmed by simulations started from initial populations in a random gametic phase, i.e. with the signs of the alleles in QTLs on a chromosome assigned randomly (all individuals remained, however, heterozygous for all marker loci and QTLs).

Figure 2 shows the dynamics of the phenotypic mean in 30 replicated runs with the BASE parameter set. The response to MAS is much stronger if the initial population is in total coupling than in total repulsion phase. The same, however, is also true for purely phenotypic selection (Fig. 4).

Figure 3 presents the efficiency in generations of MAS for the BASE parameter set. Generation 0 corresponds to the initial F1 cross, whereas generation 1 corresponds to the F2 cross (the first generation of selection). Points on the graph designated as 're-evaluation' and 'no re-evaluation' refer to simulations with and without marker re-evaluation, as explained in the Methods section. It is seen that selection is more efficient if markers are re-evaluated each generation than if they are evaluated only once. Starting from the initial population in total coupling phase, MAS with marker re-evaluation is effective (yields a higher response than purely phenotypic selection) at least until generation 20, whereas without re-evaluation it becomes ineffective after generation 9. MAS without re-evaluation started from a population in total repulsion phase becomes ineffective after generation 4, yet it continues to be effective until generation 9 if markers are re-evaluated. Similar findings were obtained in simulations with parameter sets other than BASE. Because of this, the remaining results in the paper are reported only for runs with markers re-evaluated each generation. It should be kept in mind, however, that MAS with re-evaluation requires more markers scored each generation than MAS without re-evaluation. Because of the costs associated with scoring genetic markers, it may be more economic not to re-evaluate markers each

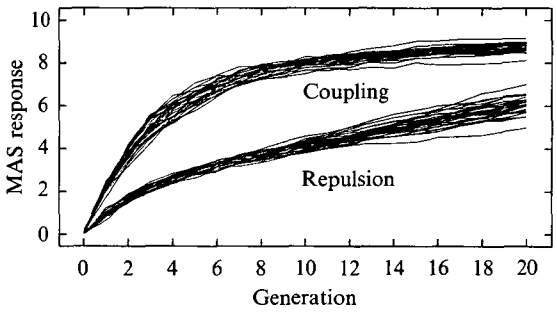

Fig. 2. Response in the phenotypic mean of one sex to MAS in 30 replicated runs with BASE parameter set.

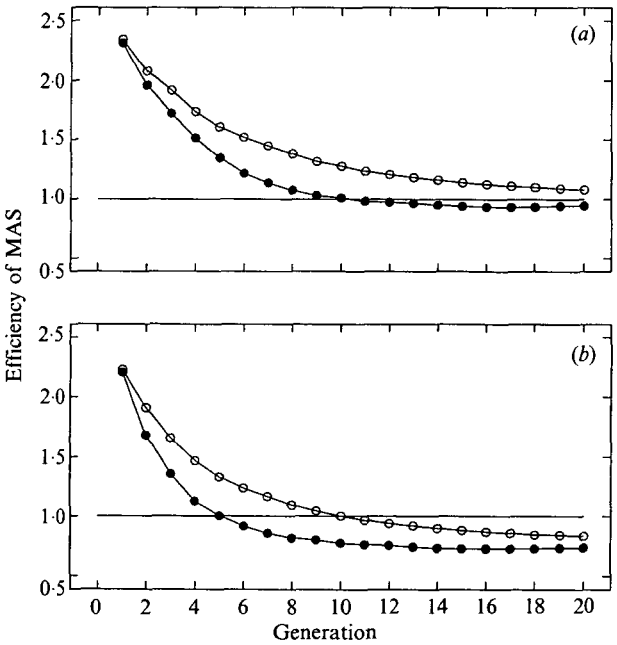

Fig. 3. Efficiency of MAS relative to purely phenotypic selection, with and without marker re-evaluation. (*a*) Coupling, (*b*) repulsion. ●, No re-evaluation; ○, re-evaluation.

https://doi.org/10.1017/S0016672300032067 Published online by Cambridge University Press

Table 2. *Efficiency of MAS*

		Markers in selection index			Markers per chromosome			Number of individuals		
Gen.	Base	3	12	20	3	6	51	100	200	3000
					Total coupling					
1	2·34	2·00	2·30	2·14	1·95	2·34	2·19	1·44	1·82	2·87
2	2·08	1·83	2·05	1·98	1·68	2·05	1·99	1·47	1·62	2·52
3	1·92	1·72	1·92	1·85	1·52	1·87	1·84	1·35	1·56	2·26
4	1·74	1·60	1·77	1·74	1·39	1·71	1·68	1·23	1·47	2·08
5	1·61	1·49	1·65	1·64	1·29	1·58	1·57	1·20	1·39	1·89
6	1·52	1·41	1·56	1·56	1·22	1·48	1·48	1·16	1·34	1·75
7	1·45	1·34	1·48	1·49	1·16	1·39	1·41	1·14	1·29	1·62
8	1·38	1·27	1·41	1·42	1·11	1·32	1·34	1·12	1·25	1·53
9	1·32	1·22	1·34	1·36	1·07	1·26	1·28	1·09	1·20	1·46
10	1·28	1·19	1·30	1·32	1·04	1·22	1·24	1·09	1·18	1·40
11	1·23	1·16	1·26	1·28	1·02	1·18	1·21	1·07	1·15	1·34
12	1·21	1·14	1·24	1·25	1·00	1·16	1·19	1·06	1·13	1·30
13	1·18	1·12	1·21	1·23	0·98	1·13	1·17	1·06	1·12	1·27
14	1·16	1·10	1·19	1·20	0·97	1·12	1·15	1·04	1·10	1·24
15	1·14	1·08	1·17	1·18	0·95	1·10	1·13	1·03	1·08	1·21
					Total Repulsion					
1	2·23	2·14	2·16	2·00	1·82	2·16	2·09	1·18	1·70	2·77
2	1·91	1·79	1·92	1·82	1·51	1·86	1·84	1·12	1·48	2·31
3	1·65	1·53	1·68	1·70	1·32	1·62	1·64	1·07	1·34	1·96
4	1·47	1·34	1·51	1·51	1·16	1·41	1·43	1·05	1·20	1·73
5	1·33	1·24	1·38	1·41	1·06	1·28	1·32	1·03	1·12	1·55
6	1·24	1·15	1·29	1·34	1·00	1·19	1·24	1·01	1·06	1·43
7	1·16	1·08	1·23	1·27	0·94	1·11	1·17	0·98	1·02	1·34
8	1·09	1·03	1·16	1·21	0·89	1·04	1·12	0·95	0·98	1·26
9	1·05	0·98	1·12	1·16	0·87	0·99	1·07	0·94	0·97	1·20
10	1·00	0·93	1·07	1·12	0·84	0·95	1·04	0·94	0·95	1·15
11	0·97	0·90	1·04	1·09	0·81	0·90	1·01	0·91	0·93	1·11
12	0·94	0·87	1·01	1·06	0·79	0·87	0·99	0·91	0·91	1·07
13	0·92	0·85	0·98	1·03	0·78	0·84	0·97	0·89	0·89	1·04
14	0·90	0·83	0·96	1·01	0·77	0·82	0·96	0·89	0·88	1·02
15	0·88	0·82	0·95	0·99	0·76	0·81	0·95	0·88	0·87	0·99

generation, even at the expense of reducing the efficiency of selection.

Table 2 and Table 3 summarize the main results of our investigation of the effects of different parameters on the efficiency of MAS. The first column in each table indicates the generation. Even though simulations were actually run for 20 generations, not much interesting information was revealed after generation 15. Therefore, only 15 generations are presented in order to save space. The second column in both tables shows the efficiency of MAS for the BASE parameter set. The numbers correspond to the points in Fig. 3 for the runs with marker re-evaluation. More parameter sets than those appearing in Tables 2 and 3 were actually investigated (see Table 1).

The results indicate that MAS can effectively utilize the linkage disequilibrium between QTLs and genetic markers created by a cross between inbred lines. For example, the response in the first generation of MAS in a population with BASE parameters is between 2·23 and 2·34 times (depending on the initial gametic phase) stronger than the response to purely phenotypic selection. The efficiency is generally higher if the initial population is in coupling gametic phase than if it is in repulsion phase. The explanation for this is that in coupling phase the contributions to the variance in the character by blocks of QTLs in linkage disequilibrium with each other are large so that genetic markers can easily detect a whole block. In the repulsion phase, however, the same blocks of QTLs contribute relatively little to the variance in the character and, consequently, they cannot be as easily detected by markers. By the time recombination separates these QTLs so they might be detected by the markers, the linkage disequilibrium between markers and QTLs may become too weak because of recombination between the markers and QTLs.

The efficiency of MAS is clearly determined not just by the number of markers included in the regression and, hence, contributing to the selection index, but, more importantly, by the significance of the effects of the markers. The more markers are in the regression, the lower is the significance of their effects. As Zhang & Smith (1992) pointed out, the inclusion of too many markers would 'add (in estimation) more noise than information to the system'. This is confirmed by

https://doi.org/10.1017/S0016672300032067 Published online by Cambridge University Press

Table 3. *Efficiency of MAS*

Gen.	BASE	Initial heritability 0·2	Initial heritability 0·4	Random mating generations 5	Random mating generations 10	Random mating generations 20	Selection strength 10%	Genetic map G25	Genetic map G13	Genetic map E10
				Total coupling						
1	2·34	1·83	1·40	1·90	1·77	1·22	2·41	2·47	2·37	2·41
2	2·08	1·65	1·32	1·79	1·55	1·14	2·00	2·07	2·12	2·03
3	1·92	1·53	1·27	1·63	1·46	1·14	1·77	1·84	1·90	1·91
4	1·74	1·43	1·21	1·55	1·38	1·10	1·60	1·71	1·75	1·77
5	1·61	1·36	1·16	1·45	1·31	1·08	1·48	1·60	1·61	1·62
6	1·52	1·28	1·13	1·37	1·26	1·06	1·37	1·51	1·50	1·53
7	1·45	1·23	1·10	1·32	1·20	1·04	1·30	1·44	1·40	1·43
8	1·38	1·18	1·07	1·26	1·15	1·02	1·24	1·38	1·33	1·36
9	1·32	1·15	1·06	1·22	1·12	0·99	1·20	1·33	1·27	1·30
10	1·28	1·12	1·05	1·17	1·09	0·96	1·17	1·30	1·22	1·25
11	1·23	1·10	1·04	1·14	1·06	0·94	1·14	1·26	1·19	1·22
12	1·21	1·09	1·03	1·12	1·04	0·93	1·12	1·23	1·16	1·18
13	1·18	1·07	1·02	1·10	1·02	0·92	1·10	1·21	1·13	1·15
14	1·16	1·06	1·02	1·09	1·01	0·92	1·09	1·20	1·11	1·12
15	1·14	1·06	1·02	1·07	0·99	0·92	1·08	1·17	1·10	1·10
				Total repulsion						
1	2·23	1·75	1·37	1·80	1·45	1·35	2·05	2·14	2·27	2·17
2	1·91	1·52	1·24	1·51	1·28	1·13	1·54	1·78	1·99	1·82
3	1·65	1·35	1·15	1·31	1·17	1·06	1·36	1·55	1·77	1·66
4	1·47	1·23	1·09	1·19	1·08	0·98	1·21	1·38	1·59	1·53
5	1·33	1·14	1·04	1·12	1·00	0·93	1·12	1·25	1·44	1·45
6	1·24	1·08	1·00	1·07	0·96	0·90	1·04	1·18	1·33	1·38
7	1·16	1·02	0·97	1·01	0·92	0·87	0·99	1·13	1·24	1·32
8	1·09	0·98	0·96	0·96	0·90	0·86	0·95	1·07	1·19	1·28
9	1·05	0·94	0·94	0·93	0·87	0·85	0·92	1·03	1·14	1·25
10	1·00	0·91	0·94	0·91	0·85	0·84	0·90	1·00	1·10	1·21
11	0·97	0·88	0·95	0·88	0·84	0·83	0·88	0·99	1·08	1·17
12	0·94	0·87	0·95	0·85	0·83	0·83	0·87	0·97	1·05	1·13
13	0·92	0·85	0·96	0·83	0·82	0·83	0·85	0·95	1·04	1·10
14	0·90	0·85	0·97	0·82	0·82	0·83	0·84	0·95	1·02	1·08
15	0·88	0·84	0·97	0·81	0·82	0·84	0·84	0·94	1·01	1·06

(Alternative genetic maps: G25: 25 QTLs with geometric series effects; G13: 13 QTLs with geometric series effects; E10: 10 QTLs with equal effects.)

columns 3–5 in Table 2 showing the effect of the number of markers included in the selection index (out of the total of 110 markers in the genome) on the efficiency of MAS. As compared to the BASE case of 6 markers in the selection index, the efficiency drops in earlier generations if the index includes 20 markers. The efficiency drops also if only 3 markers are in the selection index, although it is still surprisingly high given so few markers contribute to the index.

Having more genetic markers on a chromosome would seem to provide more opportunities for selecting markers with effects of higher significance, and, hence, to raise the efficiency of MAS. This, however, is not necessarily true, as demonstrated by columns 6–8 in Table 2. Even though the efficiency is, indeed, higher in the BASE case of 11 markers than in the case of only 3 markers per chromosome, the efficiency is not different from the BASE case if each chromosome carries only 6 markers. Moreover, with 51 markers per chromosome, the efficiency is lower than in the BASE case of 11 markers. Thus, increasing the number of markers on a chromosome does not necessarily make MAS more efficient, and may even reduce its efficiency. How many markers on a chromosome are needed in order to achieve the highest efficiency depends, of course, on the relative position of QTLs and on the population size.

Provided a population is sufficiently large, the response to purely phenotypic selection is practically not affected by its actual size, as demonstrated by the solid lines in Fig. 4 showing the response to such selection by populations of 100, 200, 500, 1000 and 3000 individuals of each sex. On the other hand, effects of the markers contributing to the molecular score become more significant with increasing population size. Consequently, MAS is expected to be more efficient in larger populations. Columns 9–11 in Table 2 as well as points plotted in Fig 4 confirm this: MAS efficiencies increase substantially in larger populations. These results as well as results for other parameter sets not reported here single out population size as the most important factor affecting the efficiency of MAS.

The efficiency of MAS declines noticeably with increasing heritability, as demonstrated by columns

https://doi.org/10.1017/S0016672300032067 Published online by Cambridge University Press

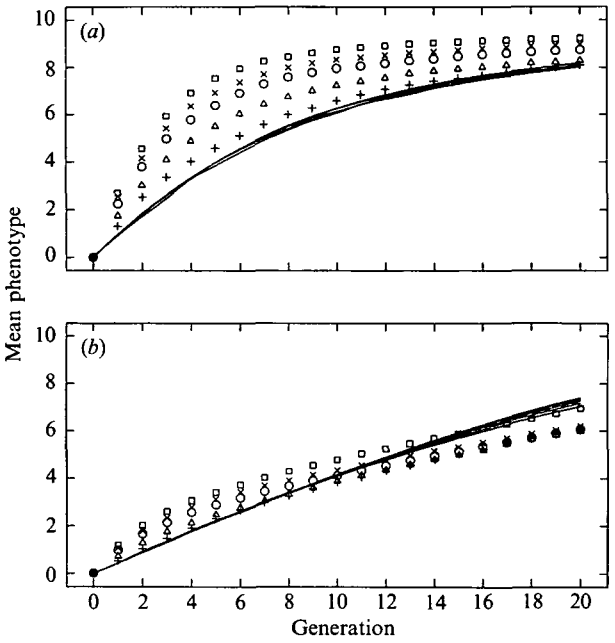

Fig. 4. Dynamics of the mean phenotype under purely phenotypic selection (solid lines) and MAS (scattered plots) in populations of different sizes. (*a*) Coupling, (*b*) repulsion. Number of individuals of each sex. □, 3000; ×, 1000; ○, 500; △, 200, +, 100.

3–4 in Table 3. This is expected since the phenotype becomes a better predictor of the individual's genotypic value when the heritability is high and, hence, less information about the genotypic value is gained by using genetic markers (Lande & Thompson, 1990; Zhang & Smith, 1992).

Given that MAS is more effective if the heritability is low and that the heritability is reduced by selection, one might think that the efficiency of MAS should increase as selection progresses. Yet, the efficiency always declines in the course of selection for any set of parameters. This is because the linkage disequilibrium between markers and QTLs is gradually destroyed by recombination. Also, the frequencies of QTLs (or blocks of linked QTLs) having large effects as well as of markers associated with them become rapidly reduced by selection.

It is important to note that crossing inbred lines creates linkage disequilibrium not only between markers and QTLs but between different markers as well. While disequilibrium between markers and QTLs is utilized by MAS, disequilibrium between markers impedes the detection of 'good' markers, i.e. those that are closest to QTLs. Allowing recombination to reshuffle markers before initiating selection weakens associations between 'good' markers and QTLs. Columns 5–7 in Table 3 show this to be true, the efficiency of MAS was reduced in populations undergoing generations of random mating before selection. This was also demonstrated by Zhang & Smith (1992).

The optimal number of markers per chromosome increases with generations of random mating before selection. For example, after 20 generations of random mating following hybridization, with QTLs initially in coupling phase and 6 markers in the index, the efficiency of MAS in the first generation of selection using 21 markers per chromosome was 1·69, as compared to the efficiencies of 1·22 using 11 markers per chromosome (Table 3 column 7), 1·64 using 51 markers per chromosome, and 1·48 using 101 markers per chromosome. Thus, the previous conclusion, that increasing the number of markers does not necessarily make MAS more efficient, still holds, even though the optimal number of markers depends on the degree of linkage disequilibrium.

Increasing selection strength (selecting 10% of individuals rather than 25%) seems to slightly reduce the efficiency of MAS, particularly if the initial population is in repulsion gametic phase (column 8 in Table 3).

Simulations were conducted with maps that differed from the one in Figure 1 by the location of QTLs in the genome, their total number and their allelic effects. The results of these simulations are presented in the last three columns of Table 3. There does not appear to be much of an effect for these maps.

The value of the coefficient b_M in the index I in equation (11) turned out to be extremely high in practically all of our simulations, so that selection was based almost exclusively on the molecular score. It is clear that if very many markers are in the regression, they may account for almost all of the genotypic variation resulting in a large value of the coefficient b_M. In our simulations, however, this appears to be true even with only a few markers in the regression. This can be due to the correlation between different markers as well as between different QTLs, so that even a few markers may account for a large proportion of the genotypic variation. That nearly all of the weight in the selection index was on the molecular score evidently results from low heritability of the character and also from large sample sizes (at least 100 individuals of each sex) so that the proportion of additive genetic variance explained by the markers, p_M, was near one. This differs from the simulations of Zhang & Smith (1992) in which selection based strictly on the molecular score was always less effective than selection based on the combined index. It should be kept in mind, however, that their index incorporated not the phenotype of an individual but rather the BLUP estimate of the individual's genotypic value using family data (Kennedy & Sorensen, 1988). The heritability of the BLUP estimate is higher than the heritability of the phenotype. Since higher heritability results in less efficient MAS, selection based exclusively on the molecular score was less efficient than selection based on their combined index and even than selection based exclusively on the BLUP estimates. It should also be noted that Zhang and Smith evaluated markers for computing molecular scores only in the first generation of a computer run. But, as our results demonstrate, the efficiency of MAS is lower if markers

https://doi.org/10.1017/S0016672300032067 Published online by Cambridge University Press

Table 4. *Markers contributing to the molecular score and their regression coefficients in 3 generations of 2 replicated runs for one sex with BASE parameter set (Coupling initial populations)*

Gen.	[Chromosome:marker] Regression coefficient					
Run 1						
1	[1:2]	[3:11]	[4:3]	[6:11]	[7:1]	[7:2]
	1·21	1·14	1·24	1·67	−2·19	4·56
2	[1:5]	[1:8]	[1:10]	[2:2]	[6:3]	[10:8]
	2·35	−2·23	1·48	−1·09	1·38	1·60
3	[1:4]	[1:7]	[6:6]	[7:1]	[7:7]	[8:3]
	1·38	0·82	−0·72	0·87	0·76	1·29
Run 2						
1	[1:4]	[2:4]	[4:7]	[7:4]	[9:8]	[10:3]
	1·91	1·19	1·07	2·01	1·48	1·27
2	[1:3]	[3:10]	[3:5]	[7:3]	[8:4]	[9:3]
	1·28	1·42	−1·06	2·16	1·09	0·91
3	[2:5]	[6:1]	[7:1]	[8:9]	[8:4]	[10:9]
	0·86	1·40	2·83	0·95	1·58	1·76

are not re-evaluated each generation. It is not only that markers contributing to the molecular score were not re-evaluated in subsequent generations of their runs, but the effects (regression coefficients) of these markers were also not re-evaluated, retaining in each generation of a run the initially obtained values. This almost certainly has lowered even further the effectiveness of MAS in the simulations by Zhang and Smith.

Table 4 shows markers that were selected by two-stage regressions as the most significant among all 110 markers in the genome of the map in Fig. 1 during three generations of 2 replicated runs with the BASE set of parameters started from initial populations in the coupling phase. It is seen that the composition of markers selected by the regression, and, hence, contributing to the molecular score changes between replicate runs, and even between generations of the same run. This is why re-evaluating markers each generation increases the efficiency of MAS. It is also seen that the significance of a marker is not necessarily a reflection of its close linkage to a QTL. Indeed, marker 11 on chromosome 6 which is selected in the first generation of run 1 is located quite far from a QTL (Fig. 1). It should also be noticed that since the initial populations of the two runs are in total coupling phase, all alleles of QTLs as well as of marker genes on a chromosome have the same sign. Consequently, the additive effects of markers located on one chromosome must be of the same sign. Yet, the regression coefficients of markers 1 and 2 on chromosome 7 selected in the first generation of run 1 have opposite signs. This may be a result of the high correlation (colinearity) of adjacent markers in generation 1, as indicated by the large magnitudes of the markers with opposite signs on chromosome 7. However, some other negative signs in Table 4 cannot be so explained, because the corresponding markers are unlinked or loosely linked to other markers included in the selection index. Hence, the regression coefficient of a marker is not exactly its 'additive effect' as implied by the notion of Marker–QTL association. Evidently, even when it is quite efficient, MAS is not necessarily utilizing stable Marker-QTL associations.

The regression methodology we employed in simulating MAS is designed to maximize the immediate response to selection in each generation, based on estimated additive effects associated with markers. This makes little use of map information for markers, other than choosing evenly spaced markers on each linkage group. Maximum likelihood interval mapping (Lander & Botstein 1988; Paterson *et al.* 1988, 1990) provides a map-based method of estimating additive effects associated with markers, but essentially the same information can be recovered from multiple regression methods (Haley & Knott 1992). Any map-based method of MAS should allow for our finding that random genetic drift sometimes causes distantly linked (or unlinked) markers to be most highly associated with a particular QTL. Other approaches to MAS, focused more directly on optimizing long-term gains, e.g. utilizing dominance and epistatic effects, might help to increase long-term selection efficiency. Nevertheless, with large population sizes, our method does show greatly increased efficiency for several generations in comparison to purely phenotypic selection.

4. Conclusions

The main conclusion of this investigation is that MAS employing multiple regression of the phenotype on genetic markers can utilize the linkage disequilibrium created in a cross of inbred lines. Such selection is more effective than conventional selection based exclusively on phenotypes of individuals, at least in

https://doi.org/10.1017/S0016672300032067 Published online by Cambridge University Press

the first several generations. The efficiency of selection is substantially higher if genetic markers contributing to the molecular score are re-evaluated each generation than if they are evaluated only once. Increasing the number of markers on a chromosome does not necessarily result in more efficient selection and too many markers may actually lower the efficiency. In our simulations, nearly all of the weight in the selection index was on the molecular score (practically pure marker selection) because of the low heritability of the character and the large population sizes. Of all the factors investigated here, population size was the most important in determining the efficiency of MAS.

We wish to thank R. Thompson and W. G. Hill for helpful discussions. A.G. would like to thank B. Charlesworth for his help. This work was supported by U.S. Public Health Service grant GM27120.

References

Draper, N. R. & Smith, H. (1981). *Applied Regression Analysis*. New York, Chichester, Toronto: Wiley.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Kennedy, B. W. & Sorensen, D. A. (1988). Properties of mixed-model methods for prediction of genetic merit. In *Quantitative Genetics* (ed. B. S. Weir, E. J. Eisen, M. M. Goodman and G. Namkoong), pp. 91–103. Massachusetts: Sinauer.

Lande, R. (1981). The minimum number of genes contributing to quantitative variation between and within populations. *Genetics* **99**, 541–553.

Lande, R. (1992). Marker-assisted selection in relation to traditional methods of plant breeding. In *Plant Breeding in the 1990s* (ed. H. T. Stalker and J. P. Murphy), pp. 437–451. Wallingford: CAB International.

Lande, R. & Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**, 743–756.

Lander, E. S. & Botstein, D. (1988). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E. & Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors using a complete linkage map of restriction fragment length polymorphisms. *Nature* **335**, 721–726.

Paterson, A. H., DeVerna, J. W., Lanini, B. & Tanksley, S. D. (1990). Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes in an interspecies cross of tomato. *Genetics* **124**, 735–742.

Zhang, W. & Smith, C. (1992). Computer simulation of marker-assisted selection utilizing linkage disequilibrium. *Theoretical and Applied Genetics* **83**, 813–820.

https://doi.org/10.1017/S0016672300032067 Published online by Cambridge University Press