Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits

Russell Lande* and Robin Thompson†

*Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, and †Institute of Animal Physiology and Genetics Research, Edinburgh Research Station, Roslin, Midlothian EH25 9PS, Scotland

Manuscript received July 17, 1989
Accepted for publication November 14, 1989

ABSTRACT

Molecular genetics can be integrated with traditional methods of artificial selection on phenotypes by applying marker-assisted selection (MAS). We derive selection indices that maximize the rate of improvement in quantitative characters under different schemes of MAS combining information on molecular genetic polymorphisms (marker loci) with data on phenotypic variation among individuals (and their relatives). We also analyze statistical limitations on the efficiency of MAS, including the detectability of associations between marker loci and quantitative trait loci, and sampling errors in estimating the weighting coefficients in the selection index. The efficiency of artificial selection can be increased substantially using MAS following hybridization of selected lines. This requires initially scoring genotypes at a few hundred molecular marker loci, as well as phenotypic traits, on a few hundred to a few thousand individuals; the number of marker loci scored can be greatly reduced in later generations. The increase in selection efficiency from the use of marker loci, and the sample sizes necessary to achieve them, depend on the genetic parameters and the selection scheme.

ARTIFICIAL selection on the phenotypes of domesticated species has been practiced consciously or unconsciously for millennia, with dramatic results. Recently, advances in molecular genetic engineering have promised to revolutionize agricultural practices. There are, however, several reasons why molecular genetics can never replace traditional methods of agricultural improvement, but instead they should be integrated to obtain the maximum improvement in the economic value of domesticated populations.

1. The rate of improvement of economically important characters such as grain yield in corn and wheat, and milk yield in dairy cattle, has been a few to several percent of the mean per year for the past several decades (SMITH 1988; FEHR 1984). For various crop plants it has been established that roughly half of this improvement is due to improved husbandry practices, i.e. environmental effects rather than genetic changes (FEHR 1984).

2. Most characters of economic importance are quantitative traits, influenced by numerous loci throughout the genome that often have individually small effects (BREESE and MATHER 1957; THODAY 1961; WRIGHT 1968 Ch. 15; LANDE 1981; EDWARDS, STUBER and WENDEL 1987; WELLER, SOLLER and BRODY 1988; SHRIMPTON and ROBERTSON 1988a, b). Genes with small effects are difficult to map precisely (SMITH 1967; SOLLER and BECKMANN 1983; LANDER and BOTSTEIN 1989), and there may be practical problems of engineering polygenic traits once the genes have been identified at the molecular level.

3. Single genes of major effect that are amenable to genetic engineering usually have deleterious pleiotropic effects. This helps to explain why evolution in natural populations usually proceeds in a Darwinian fashion, by a series of small genetic steps (FISHER 1958 pp. 41–44; WRIGHT 1968 Ch. 15; LANDE 1981, 1983). Genetic engineers lately confronted this fact in the form of low viability and fertility of mice and pigs expressing transgenic growth hormones (PURSEL *et al.* 1989). The deleterious pleiotropic effects of an engineered gene may be ameliorated by artificial and/or natural selection of polygenic modifiers (CASPARI 1952; WRIGHT 1977 p. 463).

4. The high mutability of polygenic characters guarantees genetic variation will arise within populations that can be usefully selected to improve on whatever previous gains have been made. For typical quantitative characters, in excess of one per hundred gametes contains a new mutation having a detectable effect (SPRAGUE, RUSSELL and PENNY 1960; RUSSELL, SPRAGUE and PENNY 1963; HOI-SEN 1972). New additive genetic variance arises by mutation at a rate on the order of 10^{-3} times the environmental variance per generation in characters in a variety of species (LANDE 1975; LYNCH 1988). At these rates, spontaneous mutation has been judged to be important in long-term

The publication costs of this article were partly defrayed by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

selection programs lasting more than about 20 generations (HILL 1982a, b).

One method of integrating molecular genetics with artificial selection is known as marker-assisted selection (MAS). In this paper we explore the practical utility of MAS by analyzing classical schemes of individual and family selection. Optimal selection indices combining phenotypic and molecular information are derived to maximize the rate of improvement of quantitative characters. The efficiency of these selection indices relative to purely phenotypic selection is analyzed. We also investigate the number of marker loci and the population sample sizes necessary to implement MAS.

NUMBER OF MARKER LOCI NEEDED TO DETECT QTLs

The basic theory for incorporating specific loci with direct effects on a quantitative character into a selection index was derived by NEIMANN-SORENSEN and ROBERTSON (1961) and SMITH (1967). Marker loci with no direct effect on the character(s) of interest also can be utilized in selection because of statistical associations (linkage disequilibria) between alleles at the marker loci and quantitative trait loci (QTLs) (SOLLER 1978; STUBER *et al.* 1980; STUBER, GOODMAN and MOLL 1982; TANKSLEY, MEDINA-FILHO and RICK 1981, 1982; SOLLER and BECKMANN 1983, 1988; HELENTJARIS *et al.* 1986; STAM 1986; SMITH and SIMPSON 1986; PATERSON *et al.* 1986; LANDER and BOTSTEIN 1989). A major limitation in utilizing these associations for artificial selection is that recombination in will reduce the linkage disequilibria and diminish the effectiveness of selection on the marker loci, unless the markers and the QTLs are very tightly linked.

Linkage disequilibria between pairs of loci are produced by three factors: hybridization, random genetic drift, and epistatic selection. In a large population created by hybridization between genetically differentiated groups, after T generations of random mating substantial linkage disequilibria are likely to be maintained between selectively neutral loci with recombination rates $r < 1/T$ (KIMURA and OHTA 1971). Genetic drift in a randomly mating population of effective size N_e is expected to produce substantial associations between polymorphic loci with recombination rates $r < 1/(4N_e)$ (HILL and ROBERTSON 1968). Thus, apart from the effects of selection, substantial associations between marker loci and QTLs are expected when recombination rates are less than $r^* = \max[1/T, 1/(4N_e)]$.

The minimum number of molecular markers, located randomly in the genome, needed to detect (in a large sample from the population) most of the genetic variance at "important" QTLs with appreciable effects can be determined roughly as follows. In a genome with a total recombination map length of L Morgans the average recombination rate between adjacently linked molecular markers must be at most r^* if substantial linkage disequilibria are to be expected between the markers and QTLs located randomly in the genome. Therefore, the number of molecular markers must be at least about

$$2L/r^* + C = \min[2TL, 8N_eL] + C \qquad (1)$$

where C is the haploid chromosome number. In domesticated populations, except those of very small size, the number of generations since the last hybridization event usually will be smaller than four times the effective population size, $T < 4N_e$. Hybridization is therefore generally a more powerful mechanism for generating useful linkage disequilibria than is random genetic drift. The number of marker loci necessary for the likely detection of associations with the important QTLs is thus usually about $2TL + C$. This number would be smaller for populations that are partially inbred due to subdivision or matings between close relatives; in highly self-fertilizing plant populations, $4(1 - 1/2^T)L + C$ molecular markers would be likely to detect substantial associations with important QTLs T generations after hybridization.

TABLE 1

Minimum numbers of randomly placed molecular marker loci for the likely detection of substantial associations with important QTLs at different times after hybridization in typical domestic animal and crop plant genomes (from Equation 1 and text)

	Map length (Morgans)	Haploid chromosomes	Breeding system	Generations since hybridization		
				1	5	10
Animal	30	20	Random mating	80	320	620
Plant	10	10	Random mating	30	110	210
			Selfing	30	49	50

Numbers of molecular markers required to be useful for various timespans after a hybridization event are given in Table 1 for typical domestic animal and crop plant genomes (KING 1974, 1975). These numbers neglect the possible effects of natural and artificial selection after the first generation: strong heterosis in a cross can help to maintain linkage disequilibria (LEWONTIN 1964); artificial directional selection can reduce linkage disequilibria (FELSENSTEIN 1965) and cause fixation of polymorphisms. For highly self-fertilizing plants, MAS may be useful for only a few generations following hybridizing, since selection among inbred lines can be performed accurately based solely on the phenotypic scores of many individuals per line.

IMPROVEMENT OF A SINGLE CHARACTER

Individual selection: Consider a random mating population with no sexual dimorphism, and discrete nonoverlapping generations. The additive genetic effects of QTLs associated with linked molecular markers can be estimated by multiple regression of individual phenotypic value, z, on the number of copies of a particular allele (0, 1, or 2) at the polymorphic marker loci. For molecular polymorphisms at the level of single nucleotide sites there generally will be only two alleles segregating in the population. If a linkage map of the marker loci has been constructed, a separate multiple regression can be performed for each linkage group, since in a randomly mating population there is not likely to be much linkage disequilibria between loci on different chromosomes. Multiple regression will account for linkage disequilibria among the different marker loci associated with linked QTLs, as well as linkage disequilibria among the QTLs. With a sufficient number of linked markers in a sufficiently large sample of individuals, nearly all of the additive genetic variance in the character contributed by a particular QTL can be accounted for, even if each marker separately has only an imperfect association with the QTL. Thus when multiple markers are associated with a given QTL, little extra information can be gained by interval mapping based on maximum likelihood (LANDER and BOTSTEIN 1989) in comparison with the standard methodology of multiple regression.

Classical selection theory, based on the assumption of small gene frequency changes each generation, indicates that selection on the additive effects of QTLs will maximize the current rate of response to selection within a randomly mating population (FALCONER 1981; SIMMONDS 1981). Incorporation of dominance and epistatic interactions in the selection index could be accomplished by including in the regression nonlinear terms involving products of the numbers of particular alleles at the marker loci. This would lead to serious statistical difficulty involving the estimation of many more coefficients in the selection index; the dimension of the problem can be reduced by considering nonlinear terms only for marker loci with significant additive effects. Even then there are further complications in predicting and optimizing the contribution to the long-term response made by selection directly on nonadditive genetic effects. In this paper we restrict our analysis to only the additive genetic effects of QTLs.

As selection proceeds, the associations among the marker loci and the QTLs will change due to recombination, random genetic drift, and selection; this may necessitate reevaluation of the associations every few or several generations. In the first generation following hybridization (the F_2 or backcross), a full multiple regression of phenotype on the numbers of alleles at all marker loci (on a given chromosome), will yield markers associated with the largest apparent additive effects that generally are overestimated because of sampling errors. In subsequent generations of selection before the next hybridization event, a simple procedure will provide unbiased estimates of the additive effects associated with the marker loci included in the selection index: first, choose the marker loci to be included in the selection index based on information from the multiple regression in the previous generation(s) (*e.g.*, by stepwise multiple regression); then, using information from the current generation, perform separate multiple regression(s) with only those markers chosen to be in the index. Because the marker loci included in the selection index in the current generation have been chosen *a priori* based on independent data, their estimated additive effects (partial regression coefficients) will be unbiased (KENDALL and STUART 1973). Thus, after determining the marker loci having significant associations with QTLs in the initial generation following hybridization only those markers need be scored in subsequent generations of selection until the associations are reevaluated.

Marker loci associated with highly significant additive effects on the character can be included in a net molecular score, m, which for any individual is the sum of the additive effects on the character associated with these markers. Use of the net molecular score, instead of multivariate molecular data with separate contributions from individual QTLs, is justified in APPENDIX III. The selection index

$$I = b_z z + b_m m \qquad (2)$$

is optimized by choosing the weight coefficients b_z and b_m to maximize the rate of improvement in the mean phenotype per generation, using classical theory (see APPENDIX I). Because the molecular score has no intrinsic economic value, the relative weights are found to be

$$b_m/b_z = (1/h^2 - 1)/(1 - p) \qquad (3)$$

in which h^2 is the heritability of the character (the proportion of the total phenotypic variance due to additive effects of all QTLs) and p is the proportion of the additive genetic variance in the character that is associated with the marker loci (NEIMANN-SORENSEN and ROBERTSON 1961).

The efficiency of this selection index in very large samples, compared to purely phenotypic selection, can be expressed as a ratio of the rate of response in the mean phenotype per generation under index selection to the rate of response under conventional phenotypic selection with the same intensity of selection. The latter is $h^2\sigma_z i$ in which i is the intensity of selection, or

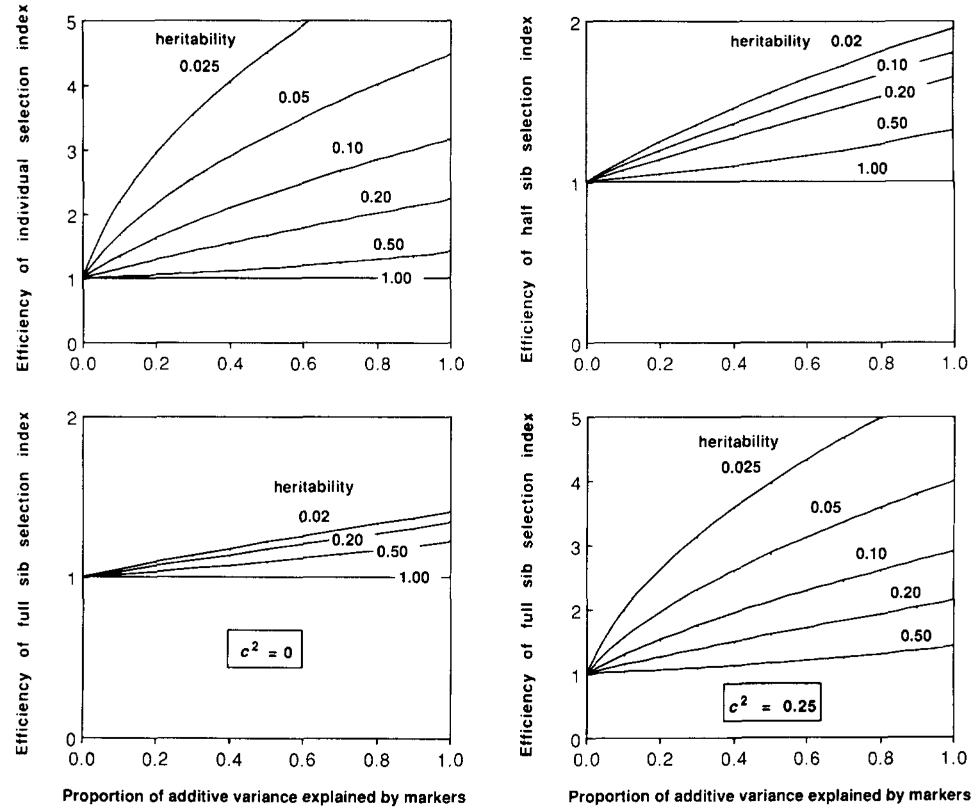

FIGURE 1.—Efficiency of marker assisted selection in the improvement of a single character, relative to traditional methods of phenotypic selection with the same selection intensity, assuming very large sample sizes. Relative efficiencies are plotted as a function of p, the proportion of the additive genetic variance in the character significantly associated with the marker loci, for various values of h^2, the heritability of the character. *Upper left:* individual selection. Other graphs depict combined selection on individual and family data, assuming very large families. *Upper right:* paternal-half sibs. *Lower left and right:* full sibs. c^2 is the fraction of the total phenotypic variance due to common family environment and genetic dominance. From Equations 4 and 9.

standardized selection differential—the difference between the mean phenotype of selected and unselected individuals divided by the phenotypic standard deviation, σ_z (FALCONER 1981 Ch. 10). Assuming that both z and I are normally distributed, a given proportion of the population saved in selection corresponds to the same intensity of selection in each case, hence

$$\text{Relative efficiency} = \sqrt{\frac{p}{h^2} + \frac{(1-p)^2}{1-h^2p}}. \quad (4)$$

The phenotypic distribution is often approximately normal, at least on a transformed scale of measurement (WRIGHT 1968 Chs. 10, 11; FALCONER 1981 Ch. 17), although m need not be normal if markers associated with only a few QTLs are included in the molecular score. Nevertheless, the selection index I will be approximately normal when z is normal since relatively large weight will be placed on m only when a large fraction of the additive genetic variance is associated with the marker loci (and hence multiple QTLs have been detected, assuming the character is polygenic).

The relative efficiency of MAS is plotted in Figure 1 as a function of p for various values of h^2. For a character with $h^2 = 1$ the phenotype of an individual is a perfect indicator of its breeding value, and no extra information is provided by the marker loci. The relative efficiency of MAS on individuals can be very large for a character with low heritability if a substantial fraction of the additive genetic variance is associated with the markers. The maximum relative efficiency of MAS on individuals, attained when all of the additive genetic variance is explained by the markers ($p = 1$) so that all of the weight in the selection index is put on the molecular score, is $1/h$.

More generally, the efficiency of selection only on the marker loci, relative to phenotypic selection of the same intensity, is $\sqrt{p/h^2}$. Thus when the proportion of the additive genetic variance explained by the marker loci exceeds the heritability of the character, selection on the marker loci alone is more efficient than selection on the individual phenotype (SMITH 1967).

Sex-limited trait: When the character is expressed only in one sex, say females, artificial selection on the index can be exerted only on that sex, but selection on the molecular score can also be practiced in the opposite sex. This leads to additional gains in the rate of response of the mean phenotype in comparison with phenotypic (sex-limited) selection. Defining the intensity of selection on the index in females as $i_♀$, and on the molecular score in males as $i_♂$, the efficiency of MAS on a sex-limited character relative to phenotypic selection on females (with selection intensity $i_♀$) is

$$\text{Relative efficiency} = \sqrt{\frac{p}{h^2} + \frac{(1-p)^2}{1-h^2p}} + \frac{i_♂}{i_♀}\sqrt{\frac{p}{h^2}}. \quad (5)$$

In this case, even for a character with high heritability, information from the marker loci may greatly increase the efficiency of selection. For a character with $h^2 = 1$ the relative efficiency of MAS on individuals for a sex-limited trait is $1 + (i_♂/i_♀)\sqrt{p}$. The maximum efficiency (when $p = 1$) for any heritability is $(1 + i_♂/i_♀)/h$, which exceeds that for a character with no sexual dimorphism by a proportion $i_♂/i_♀$.

Marker selection of immatures: A two-stage selection scheme that has received considerable attention is the selection of immature individuals (seedlings, embryos or juveniles) based on molecular marker loci, followed by conventional phenotypic selection of the surviving adults (SMITH 1967; TANKSLEY, MEDINA-FILHO and RICK 1981; SOLLER and BECKMANN 1983). This creates the possibility of exerting very strong selection on the immatures even before they develop the character on which the adults are selected. The correlated response in the average adult phenotype to marker selection on the immatures is $\sigma_m i_m$ where σ_m is the standard deviation of the molecular score in immatures and i_m is the selection intensity on it. Selection on the immatures reduces the total additive genetic variance in the adult character by a proportion $p(1 - \sigma_m^{2*}/\sigma_m^2)$, in which σ_m^{2*} is the variance in m after selection on juveniles, while leaving the environmental variance unchanged. Defining the intensity of selection on the adult phenotype as i_A, the efficiency of this two-stage MAS relative to conventional phenotypic selection on the adults with the same selection intensity, i_A, is approximately

Relative efficiency

$$= \frac{i_m}{i_A}\sqrt{\frac{p}{h^2}} + \frac{1 - p(1 - \sigma_m^{2*}/\sigma_m^2)}{\sqrt{1 - h^2p(1 - \sigma_m^{2*}/\sigma_m^2)}}. \tag{6a}$$

Strong selection on the molecular score in the immature stage reduces the response to phenotypic selection on the adults, but this is more than compensated for by the initial gains from the first stage of selection. For very strong selection on the immatures (so that less than a few percent are saved), Equation 6a can be approximated by

Relative efficiency

$$= (i_m/i_A)\sqrt{p/h^2} + (1 - p)/\sqrt{1 - h^2p}. \tag{6b}$$

If the second stage of selection is based on an index including the molecular score, rather than the adult phenotype alone, the response to selection is more complicated [see COCHRAN (1951) and BULMER (1980 Ch. 11) for the solution when both z and m are normally distributed]. However, under very strong selection on the immatures, almost all of the genetic variance in the molecular score will be exhausted at the adult stage and the relative efficiency would be near that in Equation 6b. In practice the efficiency of two-stage selection may be somewhat reduced because the additive genetic effects associated with marker loci in the immatures must be estimated from adults in the previous generation before the second stage of selection, which, along with recombination, may alter the associations between the markers and QTLs.

Information from relatives: The preferred method of artificial selection for characters with low heritability is to utilize information from relatives, as this allows a more accurate estimation of an individual's breeding value than does the phenotype of the individual alone. Classical selection schemes based on full sib or half sib families are amenable to analytical treatment in the construction of a selection index that places different weights on information between and within families (LUSH 1947; ROBERTSON 1955; FALCONER 1981; BAKER 1986). We assume that a measurement of a phenotypic character and data on molecular marker loci are available for every individual subject to selection in a randomly mating population with no sexual dimorphism. Families of size n are assumed to be composed of either full sibs with a single mother and father, or paternal half sibs with a single father but a different mother for each offspring. The genetic relationship (correlation of breeding values) among family members, r, equals 1/2 for full sibs and 1/4 for half sibs. The phenotypic correlation between family members is $t = rh^2 + c^2$ in which c^2 is the fraction of the total phenotypic variance due to genetic dominance and common family environment, e.g. nonheritable maternal effects (LUSH 1947). In most breeding designs $c^2 = 0$ for paternal half sib families.

Denote the mean phenotype of a family as z_f and the phenotypic deviation of an individual from its family mean as z_w. Similarly, write the mean molecular score of a family as m_f and the deviation of an individual's molecular score from its family mean as m_w. The components of an individual's phenotype (the family mean plus the deviation from the family mean) have equal economic value, whereas the components of the molecular score have no intrinsic economic value. The selection index combining individual and family information

$$I = b_{zf}z_f + b_{mf}m_f + b_{zw}z_w + b_{mw}m_w \tag{7}$$

is optimized by choosing the weights (the b coefficients) to maximize the rate of improvement of the mean phenotype in the population (see APPENDIX I). The relative weights can be expressed in terms of $r_n = r + (1 - r)/n$ and $t_n = t + (1 - t)/n$ which are, respectively, the expected proportions of the total additive genetic and phenotypic variances found among families of size n (FALCONER 1981 Ch. 13),

$$\begin{pmatrix} b_{zf} \\ b_{mf} \\ b_{zw} \\ b_{mw} \end{pmatrix} = \begin{pmatrix} r_nh^2(1 - p)/D_f \\ [t_n - r_nh^2]/D_f \\ (1 - r)h^2(1 - p)/D_w \\ [1 - t - (1 - r)h^2]/D_w \end{pmatrix} \tag{8}$$

TABLE 2

Maximum relative efficiency of MAS (with $p = 1$ and very large sample size), in comparison with phenotypic selection of the same intensity (from Equations 4, 5, 6b and 9)

Selection scheme	Relative efficiency
Individual	
Index including markers and phenotype (on both sexes)	$1/h$
Index on female-limited trait, markers on males[a]	$(1 + i_{♂}/i_{♀})/h$
Two stage: markers on immatures, phenotype on adults (very strong selection on immatures)[b]	$(i_m/i_A)/h$
Combined individual and family index (very large families)	
Paternal half sib	$2\sqrt{(1 - h^2/4)/(1 + 2h^2)}$
Full sib,	
No common family environment or dominance	$\sqrt{2 - h^2}$
With common family environment or dominance[c]	$(2/h)\sqrt{t(1 - t)}$

[a] $i_{♂}$ and $i_{♀}$ are standardized selection differentials on males and females.

[b] i_m and i_A are standardized selection differentials on immatures and adults.

[c] $t = h^2/2 + c^2$ in which c^2 is the fraction of phenotypic variance due to common family environment and genetic dominance.

where $D_f = t_n - r_n h^2 p$ and $D_w = 1 - t - (1 - r)h^2 p$. Although data on the molecular markers in relatives provide no extra information on an individual's molecular markers, such data do allow more accurate estimation of the breeding values of the relatives and, indirectly, the breeding value of the individual as well.

The relative efficiency of MAS using information from relatives, expressed as a proportion of the response to traditional phenotypic selection on an index using the same family structure and the same overall intensity of selection, is

$$\sqrt{\frac{\dfrac{p}{h^2} + (1 - p)^2\left[\dfrac{r_n^2}{D_f} + \dfrac{(n - 1)(1 - r)^2}{nD_w}\right]}{\dfrac{r_n^2}{t_n} + \dfrac{(n - 1)(1 - r)^2}{n(1 - t)}}}. \tag{9}$$

For a family size of one ($n = 1$ and $t_1 = r_1 = 1$) this expression reduces to that in Equation 4 for individual selection.

Table 2 gives the maximum relative efficiency of MAS for various selection schemes, when all of the additive genetic variance is explained by the molecular marker loci and sample sizes are very large. For simplicity, the formulas in the table for combined individual and family selection are based on the assumption of large families, $n \gg \max[1/r, 1/t] - 1$. This is the situation when the most information is available from relatives, and the scope for improving the efficiency of selection with additional data on molecular marker loci is minimal. The same assumptions of large families and large sample sizes were used in constructing the graphs for combined individual and family selection in Figure 1, which give the relative efficiency of MAS as a function of the proportion of the additive genetic variance explained by the marker loci, p, for various values of the heritability of the character.

Figure 1 indicates that, for characters of low heritability, MAS may be considerably more efficient than conventional selection schemes based on combined individual and family information, if a substantial fraction of the additive genetic variance in the character is associated with the marker loci. For characters of low heritability, MAS using large paternal half sib families is at most twice as efficient as conventional phenotypic selection on combined individual and half sib family data. The maximum relative efficiency of MAS on full sib families for characters of low heritability with no common family environment or genetic dominance is only $\sqrt{2} = 1.41$, because the genetic relationship among full sibs is higher than that of half sibs. However, if the proportion of the total phenotypic variance due to common family environment or genetic dominance is substantial, the relative efficiency of MAS for combined individual and full sib selection may be quite high.

Figure 2 shows how family size influences the efficiency of MAS when $p = 1$ in comparison to purely phenotypic selection with the same family structure and overall selection intensity. For characters of low heritability, with no genetic dominance or common family environment, it can be seen that family sizes must be very large before the relative efficiency of MAS is as low as that on the corresponding graphs in Figure 1 at $p = 1$. However, for full sib families with appreciable variance due to genetic dominance or common family environment, the condition for large families ($n \gg 1/t - 1$) indicates that the values in the lower right graph of Figure 1 would be approached more rapidly.

Additive genetic variance explained by marker loci: We now wish to gain some idea of the proportion of the additive genetic variance in a character likely to have statistically significant association with marker loci, p. Two factors limit the statistical detectability of additive effects at QTLs: the number of marker loci,

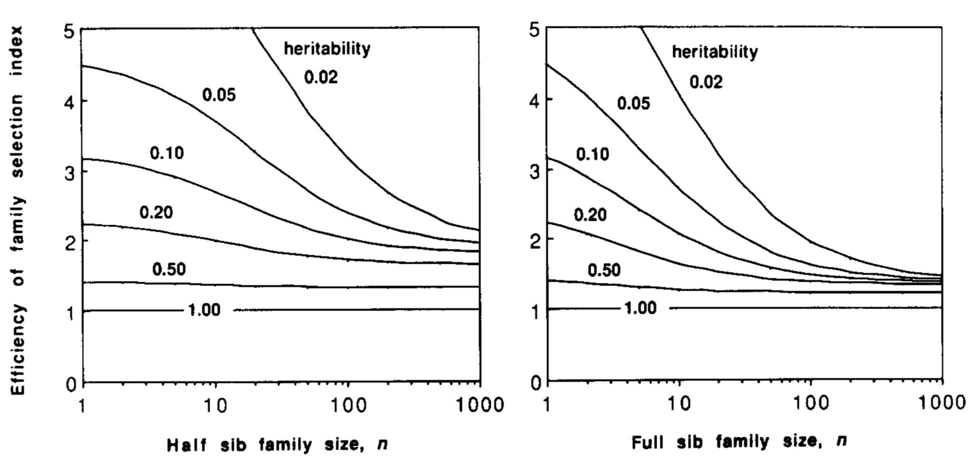

FIGURE 2.—The influence of family size, n, on the maximum efficiency of marker assisted selection, assuming that all of the additive genetic variance is explained by the marker loci ($p = 1$). The efficiency of MAS is expressed relative to the optimal scheme of combined individual and family selection only on the phenotype, from Equation 9. For full sib families it is assumed that there is no genetic dominance or common family environment ($c^2 = 0$). Values on the left side of each graph (at $n = 1$) represent individual selection; values on the right side (at large n) approach those for selection on very large families (with $c^2 = 0$ and $p = 1$) as shown in Figure 1.

and the sample size of the population in which the associations are estimated. We concluded above that hybridization is a more powerful mechanism than random genetic drift to generate linkage disequilibria between QTLs and marker loci, and that to be useful for several generations of selection the number of marker loci must be on the order of a few hundred (Table 1). Therefore we assume in this section that a very large number of markers in linkage disequilibria with the QTLs are available, and investigate only the limitation of population sample size. The value of p detected in a sample of a given size then depends primarily on the distribution of gene effects at the QTLs, since a QTL with a small effect is unlikely to be detected as statistically significant.

Most quantitative traits are influenced by numerous genes (WRIGHT 1968 Ch. 15; LANDE 1981), and typically a few loci have relatively large effects with many others having smaller effects (SPICKETT and THODAY 1966; GREGORY 1965, 1966; THOMPSON 1975; EDWARDS, STUBER and WENDEL 1987; PATERSON *et al.* 1988; SHRIMPTON and ROBERTSON 1988b). This suggests that the distribution of additive genetic variances contributed by QTLs may often be approximated by a geometric series,

$$\sigma_g^2(1 - a)[1, a, a^2, a^3, \ldots] \qquad (10)$$

which sums to the total additive genetic variance, σ_g^2, if the QTLs are in linkage equilibrium with each other. The constant a determines the relative magnitude of the contributions of each QTL. Assuming that the QTLs with substantial (detectable) effects are unlinked or loosely linked, one or two generations of recombination will suffice to bring them close to linkage equilibrium among each other, although each QTL may remain closely associated with tightly linked marker loci for several generations. (In a standard genetic cross, unlinked loci achieve linkage equilibrium in the F_2 because of nonrandom mating among the parental lines, in contrast to the asymptotic decay of linkage disequilibrium in a random mating population.) The exact form of this distribution generally will not be preserved under selection and mutation, but this model should still give a reasonable indication of the maximum proportion of additive genetic variance that can be detected with a large number of marker loci in a sample of a given size.

An informative statistic describing the evenness of the contributions of the QTLs to the total additive genetic variance is the effective number of loci. WRIGHT (1968 Ch. 15) defined such a quantity for a cross between two inbred lines as the number of unlinked loci of equal, completely additive effect that would produce the same additive genetic variance in the F_2 as in the actual cross. A different but related measure can also be defined within a single random mating population (LANDE 1981), and evaluated for the geometric series of variance contributions in formula 10 as

$$n_E = \left(\sum_{i=0}^{\infty} a^i\right)^2 \Big/ \sum_{i=0}^{\infty} a^{2i} = (1 + a)/(1 - a). \qquad (11)$$

Values of a equal to 0, 1/3, 2/3, 5/6, and 11/12 correspond to effective numbers of loci equal to 1, 2, 5, 11, and 23.

If the additive genetic variance at the lth locus in the series, $\sigma_g^2(1 - a)a^{l-1}$, is the smallest value likely to be detected in a given sample, then the maximum proportion of additive genetic variance in the character likely to be detected is

$$p = 1 - a^l. \qquad (12)$$

An estimate of the additive genetic variance contributed by a particular QTL (or linkage group) in a large sample is likely to achieve the γ level of significance if the expected value of the estimate exceeds x_γ times its standard error, where the integral of the standard normal distribution from x_γ to ∞ equals γ. Letting the

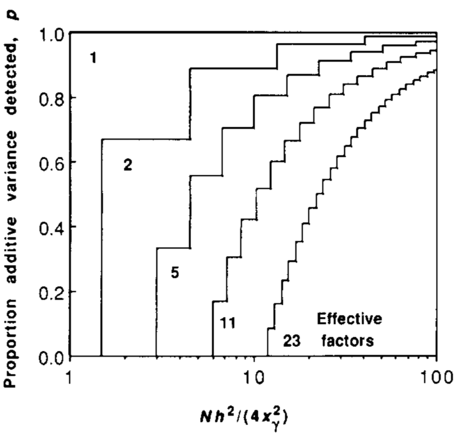

FIGURE 3.—The proportion of additive genetic variance, p, likely to be detected with a very large number of molecular marker loci, plotted against the total number of individuals in the sample, N, times $h^2/(4x_\gamma^2)$, for various values of n_E, the effective number of QTLs. x_γ is the number of standard deviations above the mean of a normal distribution needed to achieve the γ level of significance (*e.g.*, $x_{0.01} = 2.33$). Detectable QTLs are assumed to be unlinked or loosely linked and to contribute to the additive genetic variance in a geometric series. The graph is for a sample of unrelated individuals; if the sample consists of large families, its total size would have to be approximately $(1 - t)/(1 - r)$ times as large to achieve the same p value, in which t and r are the phenotypic and additive genetic correlations between family members. From Equations 10–13.

proportion of the additive genetic variance due to the ith QTL be p_i, we assume that each QTL contributes only a small fraction of the total phenotypic variance in the character, $h^2p_i \ll 1$. The statistical analysis of NEIMANN-SORENSEN and ROBERTSON (1961) and SMITH (1967) is generalized in APPENDIX II to include multiple marker loci associated with a particular QTL. The approximate results are as follows.

In a sample of N individuals that are unrelated (or not closely related) the ith QTL is likely to be detected (with up to several significantly associated markers) if

$$p_i > 4x_\gamma^2/h^2N. \tag{13a}$$

In a sample totalling N individuals grouped into families of large size, the ith QTL is likely to be detected (with up to several significantly associated markers) if

$$p_i > 4x_\gamma^2(1 - t)/(1 - r)h^2N \tag{13b}$$

in which t and r are the phenotypic and additive genetic correlations between family members. Equation 12 can be used with 13a or 13b to calculate the maximum proportion of additive genetic variance likely to be detected in a sample of N individuals, as shown in Figure 3.

It can be noticed from Equations 13 and Figure 3 that the product h^2N plays a crucial role in determining the magnitude of additive genetic variance at any QTL that can be detected as statistically significant. For a given sample size, there is an inverse relationship between the heritability of the character and the proportion of additive genetic variance that can be detected, even with a very large number of marker loci. Thus the high relative efficiencies of MAS that appear in Figure 1 for characters of low heritability with intermediate or large p values are unlikely to be realized unless sample sizes are quite large.

The total number of individuals needed to detect a given proportion of the additive genetic variance associated with the marker loci in a population structured into large families is approximately $(1 - t)/(1 - r)$ times that in a population of unrelated (or distantly related) individuals. This factor is less than one when $c^2 > r(1 - h^2)$, but is greater than one otherwise. Thus, for detecting additive genetic variance in characters of low heritability, a population composed of large families is less efficient than a population of the same size composed of unrelated individuals, unless a large fraction of the phenotypic variance is caused by genetic dominance and common family environment.

The maximum relative efficiency of MAS on unrelated individuals that can be realized for various sample sizes is shown in Figure 4, using for example the $\gamma = 0.01$ level of significance for detection of a QTL. The lines in the figure intersect because for characters with different heritabilities, but the same n_E, a larger sample size is needed to detect a given proportion of additive genetic variance in a character of lower h^2, and for a given p value the relative efficiency of MAS is larger for traits with lower h^2. From Figure 4 it is also apparent that, for a character with a given heritability, the larger the effective number of QTLs the larger is the sample size needed to obtain a given p value, because the additive variance contributed per locus decreases as n_E increases.

Loss of efficiency from sampling error: Another practical limitation on the relative efficiency of MAS occurs because the relative weights on the phenotypic and molecular information in the selection index are derived from estimated parameters. Sampling errors in the parameter estimates reduce the efficiency of MAS by causing the weight coefficients to deviate from their optimal values. The two-stage selection procedure of Equations 6a and 6b does not rely on an index, and is not subject to this limitation. For simplicity, we assume that the standard quantitative genetic parameters are known exactly, and we consider only the sampling errors from estimation of the additive effects associated with the marker loci chosen *a priori* to be in the selection index. With individual selection based on the index in Equations 2 and 3, the expected proportional loss of efficiency in a sample of N individuals (analyzed in APPENDIX III) is approximately

$$\frac{(2h^2p + K/N)p(1 - h^2)^2}{Nh^2(1 - ph^2)[p + h^2(1 - 2p)]^2}. \tag{14}$$

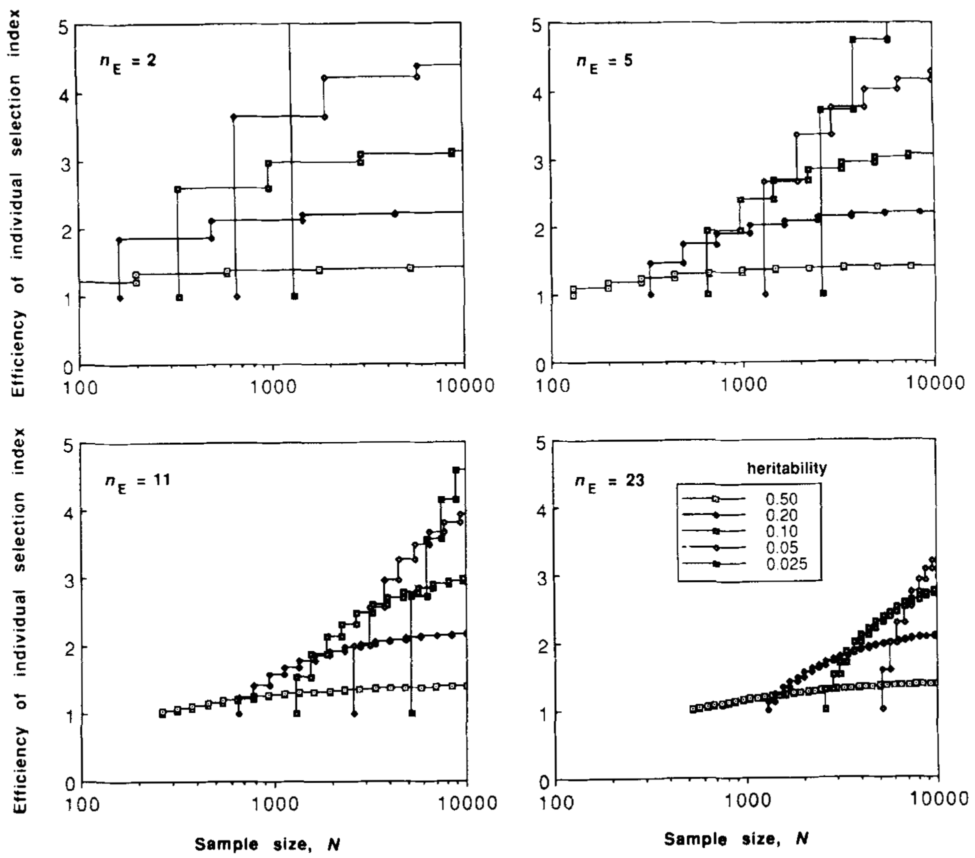

FIGURE 4.—Maximum relative efficiencies of MAS that can be realized by individual selection as a function of sample size, N, for different effective numbers of QTLs, n_E, contributing to the additive genetic variance. Values of heritability associated with each line are given in the legend in the lower right graph. From Equations 4 and 10–13, using the γ = 0.01 level of significance for detection of additive genetic variance at any QTL.

where K is the total number of marker loci used in the selection index. Some numerical examples are shown in Table 3. The expected loss of efficiency would be larger in the initial F_2 or backcross generation because of bias in overestimating additive effects of QTLs most significantly associated with the markers when no prior information is available.

IMPROVEMENT OF MULTIPLE CHARACTERS

If very large samples are available, MAS on multiple traits is more efficient in a multivariate context than in a univariate analysis of total economic value alone. This is because the molecular marker loci provide different amounts of information on different characters, which affects their weightings in a multivariate selection index. We can demonstrate this most readily by generalizing the index for individual selection in Equations 2 and 3. Let **z** be a vector of quantitative traits with phenotypic and additive genetic variance-covariance matrices **P** and **G.** Define a vector of corresponding molecular scores obtained by summing the vectors of effects on the characters produced by associated molecular marker loci, **m,** having variance-covariance matrix **M.** The vector of relative economic weights of the quantitative traits is **d,** and that for the molecular markers is **0.** Using the general theory in APPENDIX I, the weight vectors $\mathbf{b}_z^T$ and $\mathbf{b}_m^T$ that maximize the rate of economic improvement in the index $\mathbf{b}_z^T\mathbf{z} + \mathbf{b}_m^T\mathbf{m}$ are given by

$$\begin{pmatrix} \mathbf{b}_z \\ \mathbf{b}_m \end{pmatrix} = \begin{pmatrix} (\mathbf{P} - \mathbf{M})^{-1}(\mathbf{G} - \mathbf{M})\mathbf{d} \\ [\mathbf{I} - (\mathbf{P} - \mathbf{M})^{-1} (\mathbf{G} - \mathbf{M})]\mathbf{d} \end{pmatrix} \qquad (15)$$

in which **I** is the identity matrix. It can be seen that the relative weights on the quantitative characters, $\mathbf{b}_z$, differ from those under purely phenotypic selection, $\mathbf{P}^{-1}\mathbf{Gd}$, and that the relative weights on the molecular scores, $\mathbf{b}_m$, are not proportional to simple economic values of the corresponding characters, **d.**

DISCUSSION

Deterministic analysis, assuming very large sample sizes, indicates that molecular marker loci can be used to substantially increase the rate of improvement in quantitative characters by artificial selection. The potential efficiency of marker assisted selection on a single trait utilizing a combination of molecular and phenotypic information, relative to standard methods of phenotypic selection, depends on the heritability of the character, the proportion of the additive genetic variance associated with the marker loci, and the selection scheme. Under individual selection, the relative efficiency of MAS is greatest for characters with low heritability, if a moderate or large fraction of the additive genetic variance is significantly associated

TABLE 3

Percent loss of efficiency in MAS expected from sampling errors in estimation of the additive genetic variance associated with marker loci chosen a priori to be in the selection index in Equations 2 and 3

h^2	N	p	(n_E, l)	Percent loss of efficiency[a]
0.1	500	0.67	(2, 1)	0.41
		0	(5, 0)	
	1000	0.89	(2, 2)	0.22
		0.55	(5, 2)	0.19
		0	(11, 0)	
	5000	0.96	(2, 3)	0.044
		0.91	(5, 6)	0.044
		0.77	(11, 8)	0.042
		0.50	(23, 8)	0.036
0.5	100	0.67	(2, 1)	1.4
		0	(5, 0)	
	200	0.89	(2, 2)	1.5
		0.55	(5, 2)	0.45
		0	(11, 0)	
	1000	0.96	(2, 3)	0.36
		0.91	(5, 6)	0.31
		0.77	(11, 8)	0.20
		0.50	(23, 8)	0.071

[a] h^2 is the heritability of the character. N is the number of individuals in the sample. For a given effective number of QTLs influencing the character, n_E, the proportion of additive genetic variance explained by the markers, p, is obtained from Equations 10–13a with a $\gamma = 0.01$ significance level for detection of a QTL (see Figure 3). The percent loss of efficiency is calculated from 100 times Equation 14, assuming that 4 marker loci per QTL are used to estimate the additive genetic variance at each of l QTLs detected, so that $K = 4l$.

with the marker loci (Equation 4; Figure 1). Further increases in the relative efficiency of MAS are possible when individuals that do not express the phenotypic traits of interest can be selected on the basis of their molecular markers, as in the case of sex-limited traits, or when additional selection can be exerted on juveniles before development of the adult phenotype (Equations 5, 6; Table 2).

The classical methods of phenotypic selection for characters with low heritability employ information from relatives to construct a selection index based on a combination of individual and family merit. The use of phenotypic information from relatives reduces the relative efficiency of MAS, but the amount of reduction depends on family size. Unless family sizes are very large there is still opportunity for substantial increase in the efficiency of selection through the use of molecular markers. Even with large families, the relative efficiency of MAS may be great if there are common family environmental effects, e.g. strong maternal effects on full sib families (Equation 9; Figures 1, 2; Table 2).

There are three practical considerations that limit the potential utility of MAS in applied breeding programs: (1) the number of molecular marker loci necessary for the existence of significant associations (linkage disequilibria) with the QTLs, (2) sample sizes needed to detect QTLs for traits with low heritability, and (3) sampling errors in the estimation of relative weights in the selection index combining molecular and phenotypic information. We discuss each of these in turn.

Molecular marker loci (protein or DNA polymorphisms located randomly in the genome) can rarely be expected to have major effects on characters of economic importance. MAS must therefore rely on linkage disequilibria between the marker loci and the QTLs, which are continually eroded by recombination. The most potent mechanism for generating linkage disequilibria is occasional hybridization between genetically differentiated lines. In the development of new varieties, especially in plants, breeders routinely cross different preexisting elite varieties to start a new cycle of selection (SIMMONDS 1981; FALCONER 1981). Such cycles of hybridization and selection are well suited to the application of MAS. For typical outcrossing species with a recombination map length of 10 to 30 Morgans, the utilization of molecular information for several generations of selection requires scoring a few hundred marker loci in the initial generation following hybridization, although fewer markers could be used in selfing species (Table 1). The number of marker loci scored in subsequent generations of selection, until the next hybridization or general reevaluation of all the markers, can be greatly reduced by neglecting those that initially were not significantly associated with the QTLs.

The larger the number of individuals sampled in a population the higher is the proportion of additive genetic variance in a character likely to be detected through associations with the marker loci. The proportion of additive genetic variance in a trait that can be detected in a sample of a given size depends on the distribution of the additive variance contributed by the individual QTLs and on whether the population consists of unrelated (distantly related) individuals or has a family structure. Rather large samples may be necessary to detect any additive genetic variance associated with marker loci in characters with low heritability. The number of unrelated individuals needed to detect substantial additive genetic variance associated with the marker loci range from a few hundred to a few thousand, depending on the heritability of the character and the effective number of QTLs contributing to the additive genetic variance. Somewhat larger sample sizes may be required for populations structured into full or half sib families, unless a large fraction of the phenotypic variance is caused by genetic dominance and common family environment (Equations 10–13; Figures 3, 4).

The loss of efficiency due to sampling errors in estimating the relative weights placed on molecular and phenotypic information in a selection index depends on the sample size used to estimate the parameters. For simplicity we assumed that the standard quantitative genetic parameters were known exactly and analyzed the influence of sampling errors only in the estimate of additive genetic variance associated with molecular markers chosen *a priori* to be in the selection index (after the initial F_2 or backcross generation). With sample sizes of a few hundred to a few thousand individuals, the expected loss of efficiency in MAS among individuals on a single character is quite small, about 1% or less (Equation 14; Table 3). Sampling errors in standard quantitative genetic parameters, such as heritabilities and genetic and phenotypic correlations between sibs, are also quite small with such sample sizes, and are not expected to cause appreciable loss of efficiency in phenotypic selection, unless multiple characters are included in the selection index (SALES and HILL 1976; HILL and THOMPSON 1978; HAYES and HILL 1981).

Many details regarding the optimal use of MAS in long-term selection programs remain to be determined by further theory and experiments. Our results are encouraging and support the conclusion that molecular genetic polymorphisms can be used to achieve substantial increases in the efficiency of artificial selection. Although the scale of this endeavor may exceed the current capability of most molecular genetic laboratories, we anticipate that improved technology in the near future will make these procedures feasible.

We thank LAURENS METS, GERARD RIEDEL, CHARLES SMITH and WILLIAM G. HILL for stimulating discussions. CHARLES SMITH and two reviewers provided constructive criticisms of the manuscript. Supported by U. S. Public Health Service grant GM27120.

LITERATURE CITED

BAKER, R. J., 1986 *Selection Indices in Plant Breeding.* CRC Press, Boca Raton, Florida.

BREESE, E. L., and K. MATHER, 1957 The organisation of polygenic activity within a chromosome in *Drosophila*. I. Hair characters. Heredity **11:** 373–395.

BULMER, M.G., 1980 *The Mathematical Theory of Quantitative Genetics.* Oxford University Press, Oxford.

CASPARI, E., 1952 Pleiotropic gene action. Evolution **6:** 1–18.

COCHRAN, W. G., 1951 Improvement by means of selection. Proc. Berkeley Symp. Math. Stat. Probab. **2:** 449–470.

EDWARDS, M. D., C. W. STUBER and J. F. WENDEL, 1987 Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics **116:** 113–125.

FALCONER, D. S., 1981 *Introduction to Quantitative Genetics*, Ed. 2. Longman, New York.

FEHR, W. R. (ed.), 1984 *Genetic Contributions to Yield Gains of Five Major Crop Plants*, Special Publ. No. 7. Crop Science Society of America, Madison, Wisc.

FELSENSTEIN, J., 1965 The effect of linkage on directional selection. Genetics **52:** 349–363.

FISHER, R. A., 1958 *The Genetical Theory of Natural Selection*, Ed. 2. Dover, New York.

GREGORY, W. C., 1965 Mutation frequency, magnitude of change and the probability of improvement in adaptation. Radiat. Bot. **5** (Supp.): 429–441.

GREGORY, W. C., 1966 Mutation breeding, pp. 189–218, in *Plant Breeding*, edited by K. J. FREY. Iowa State University Press, Ames.

HARRIS, D. L., 1964 Expected and predicted progress from index selection involving estimates of population parameters. Biometrics **20:** 46–72.

HAYES, J. F., and W. G. HILL, 1981 Modification of estimates of parameters in the construction of genetic selection indices ("bending"). Biometrics **37:** 483–493.

HAZEL, L. N., 1943 The genetic basis for constructing selection indices. Genetics **38:** 476–490.

HELENTJARIS, R., M. SLOCUM, S. WRIGHT, A. SHAEFFER and J. NIENHUIS, 1986 Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor. Appl. Genet. **72:** 761–769.

HILL, W. G., 1982a Predictions of response to artificial selection from new mutations. Genet. Res. **40:** 255–278.

HILL, W. G., 1982b Rates of change in quantitative traits from fixation of new mutations. Proc. Natl. Acad. Sci. USA **79:** 142–145.

HILL, W. G., and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38:** 226–231.

HILL, W. G., and R. THOMPSON, 1978 Probabilities of non-positive definite between-group or genetic covariance matrices. Biometrics **34:** 429–439.

HOI-SEN, Y., 1972 Is sub-line differentiation a continuing process in inbred strains of mice? Genet. Res. **19:** 53–59.

KENDALL, M. G., and A. STUART, 1973 *The Advanced Theory of Statistics. Vol. 2. Inference and Relationship*, Ed. 3. Hafner, New York.

KIMURA, M., and T. OHTA, 1971 *Theoretical Aspects of Population Genetics.* Princeton University Press, Princeton, N.J.

KING, R. C. (ed.), 1974 *Handbook of Genetics, Vol. 2. Plants, Plant Viruses, and Protists.* Plenum Press, New York.

KING, R. C. (ed.), 1975 *Handbook of Genetics. Vol. 4. Vertebrates of Genetic Interest.* Plenum Press, New York.

LANDE, R., 1975 The maintenance of genetic variability by mutation in a quantitative character with linked loci. Genet. Res. **26:** 221–235.

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

LANDE, R., 1983 The response to selection on major and minor mutations affecting a metrical trait. Heredity **50:** 47–65.

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General considerations. Heterotic models. Genetics **49:** 49–67.

LUSH, J. L., 1947 Family merit and individual merit as bases for selection. Am. Nat. **81:** 241–261, 362–379.

LYNCH, M., 1988 The rate of polygenic mutation. Genet. Res. **51:** 137–148.

NEIMANN-SORENSEN, A., and A. ROBERTSON, 1961 The association between blood groups and several production characters in three Danish cattle breeds. Acta Agric. Scand. **11:** 163–196.

PATERSON, A. H., E. S. LANDER, J. D. HEWITT, S. PETERSON, S. E. LINCOLN and S. D. TANKSLEY, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphism. Nature **335:** 721–726.

PURSEL, V. G., C. A. PINKERT, K. F. MILLER, D. J. BOLT, R. G. CAMPBELL, R. D. PALMITER, R. L. BRINSTER and R. E. HAMMER, 1989 Genetic engineering of livestock. Science **244:** 1281–1288.

ROBERTSON, A., 1955 Prediction equations in quantitative genetics. Biometrics **11:** 95–98.

RUSSELL, W. A., G. F. SPRAGUE and L. H. PENNY, 1963 Mutations affecting quantitative characters in long-time inbred lines of maize. Crop Sci. **3:** 175–178.

SALES, J., and W. G. HILL, 1976 Effect of sampling errors on efficiency of selection indices. Anim. Prod. **22:** 1–17.

SEARLE, S. R., 1971 *Linear Models.* Wiley, New York.

SHRIMPTON, A. E., and A. ROBERTSON, 1988a The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster.* I. Allocation of third chromosome sternopleural bristle effects to chromosome sections. Genetics **118:** 437–443.

SHRIMPTON, A. E., and A. ROBERTSON, 1988b The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster.* II. Distribution of third chromosome bristle effects within chromosome sections. Genetics **118:** 445–459.

SIMMONDS, N. W., 1981 *Principles of Crop Improvement.* Longman, London.

SMITH, H. F., 1936 A discriminant function for plant selection. Ann. Eugenics **7:** 240–250.

SMITH, C., 1967 Improvement of metric traits through specific genetic loci. Anim. Prod. **9:** 349–358.

SMITH, C., 1988 Potential for animal breeding, current and future, pp. 150–160 in *Proceedings of the Second International Conference on Quantitative Genetics,* edited by B. S. WEIR, E. J. EISEN, M. M. GOODMAN and G. NAMKOONG. Sinauer, Sunderland, Mass.

SMITH, C., and S. P. SIMPSON, 1986 The use of genetic polymorphisms in livestock improvement. J. Anim. Breed. Genet. **103:** 205–217.

SOLLER, M., 1978 The use of loci associated with quantitative effects in dairy cattle improvement. Anim. Prod. **27:** 133–139.

SOLLER, M., and J. S. BECKMANN, 1983 Genetic polymorphism in varietal identification and genetic improvement. Theor. Appl. Genet. **67:** 25–33.

SOLLER, M., and J. S. BECKMANN, 1988 Genomic genetics and the utilization for breeding purposes of genetic variation between populations, pp. 161–188 in *Proceedings of the Second International Conference on Quantitative Genetics,* edited by B. S. WEIR, E. J. EISEN, M. M. GOODMAN and G. NAMKOONG. Sinauer, Sunderland, Mass.

SPICKETT, S. G., and J. M. THODAY, 1966 Regular responses to selection. 3. Interaction between located polygenes. Genet. Res. **7:** 96–121.

SPRAGUE, G. F., W. A. RUSSELL and L. H. PENNY, 1960 Mutations affecting quantitative traits in the selfed progeny of doubled monoploid maize stocks. Genetics **45:** 855–866.

STAM, P., 1986 The use of marker loci in selection for quantitative characters, pp. 170–182 in *Exploiting New Technologies in Livestock Improvement: Animal Breeding,* edited by C. SMITH, J. W. B. KING and J. MCKAY. Oxford University Press, Oxford.

STUBER, C. W., R. H. MOLL, M. M. GOODMAN, H. E. SCHAFFER and B. S. WEIR, 1980 Allozyme frequency changes associated with selection for increased grain yield in maize (*Zea mays* L.). Genetics **95:** 225–236.

STUBER, C. W., M. M. GOODMAN and R. H. MOLL, 1982 Improvement of yield and ear number resulting from selection at allozyme loci in a maize population. Crop Sci. **22:** 737–740.

TANKSLEY, S. D., H. MEDINO-FILHO and C. M. RICK, 1981 The effect of isozyme selection on metric characters in an interspecific backcross of tomato: Basis of an early screening procedure. Theor. Appl. Genet. **60:** 291–296.

TANKSLEY, S. D., H. MEDINO-FILHO and C. M. RICK, 1982 Use of naturally occurring enzyme variation to detect and map genes controlling quantitative traits in an interspecific backcross of tomato. Heredity **49:** 11–25.

THODAY, J. M., 1961 Location of polygenes. Nature **191:** 368–370.

THOMPSON, J. N., 1975 Quantitative variation and gene number. Nature **258:** 665–668.

WELLER, J. I., M. SOLLER and T. BRODY, 1988 Linkage analysis of quantitative traits in an interspecific cross of tomato (*Lycopersicon esculentum* × *Lycopersicon pimpinellifolium*) by means of genetic markers. Genetics **118:** 329–339.

WRIGHT, S., 1968 *Evolution and the Genetics of Populations. Vol. 1. Genetic and Biometric Foundations.* University of Chicago Press, Chicago.

WRIGHT, S., 1977 *Evolution and the Genetics of Populations. Vol. 3. Experimental Results and Evolutionary Deductions.* University of Chicago Press, Chicago.

Communicating editor: B. S. WEIR

APPENDIX I

Define a column vector of quantitative traits, $\mathbf{z}$, with relative economic weights $\mathbf{d}$, and the linear selection index $I = \mathbf{b}^{\mathrm{T}}\mathbf{z}$ in which $\mathbf{b}^{\mathrm{T}}$ is the transpose of $\mathbf{b}$, the column vector of weight coefficients. Assuming linearity of the regression of breeding value for economic merit on the index I, the vector of weight coefficients that maximize the rate of improvement in the economic value of the population is

$$\mathbf{b} = \mathbf{P}^{-1}\mathbf{G}\mathbf{d} \tag{A1}$$

where $\mathbf{P}$ and $\mathbf{G}$ are respectively the phenotypic and additive genetic variance-covariance matrices of the traits (SMITH 1936; HAZEL 1943; FALCONER 1981). The rate of improvement per generation in the average economic value in the population in response to selection on this index is

$$\mathbf{d}^{\mathrm{T}}\Delta\bar{\mathbf{z}} = i\,\sqrt{\mathbf{d}^{\mathrm{T}}\mathbf{G}\mathbf{b}} \tag{A2}$$

in which i is the intensity of selection on the index, measured by the standardized selection differential (FALCONER 1981 Ch. 19).

For MAS on a single character in individuals, the components of $\mathbf{z}$ are the individual phenotype, z, and the molecular score, m, so that $\mathbf{z}^{\mathrm{T}} = (z, m)$. These have relative economic values $\mathbf{d}^{\mathrm{T}} = (1, 0)$ with

$$\mathbf{P} = \begin{pmatrix} \sigma_z^2 & \sigma_m^2 \\ \sigma_m^2 & \sigma_m^2 \end{pmatrix} \quad \text{and} \quad \mathbf{G} = \begin{pmatrix} \sigma_g^2 & \sigma_m^2 \\ \sigma_m^2 & \sigma_m^2 \end{pmatrix} \tag{A3}$$

where σ_z^2 and σ_g^2 are the phenotypic and additive genetic variances in the character and σ_m^2 is the additive genetic variance explained by the marker loci. Evaluation of Equations A1 and A2 using these formulas yields text Equations 3 and 4.

For MAS using information from relatives, it is convenient to work in terms of family means and individual deviations from the family means, respectively denoted by subscripts f and w for both the phenotype and the molecular index, since these quantities are uncorrelated. Thus $\mathbf{z}^{\mathrm{T}} = (z_f, m_f, z_w, m_w)$ and $\mathbf{d}^{\mathrm{T}} = (1, 0, 1, 0)$. For families of size n, the phenotypic variance-covariance matrix can be expressed in terms of the phenotypic and additive genetic correlations between relatives, t_n and r_n, defined in the text,

$$\mathbf{P} = \begin{pmatrix} t_n\sigma_z^2 & r_n\sigma_m^2 & 0 & 0 \\ r_n\sigma_m^2 & r_n\sigma_m^2 & 0 & 0 \\ 0 & 0 & (1 - t_n)\sigma_z^2 & (1 - r_n)\sigma_m^2 \\ 0 & 0 & (1 - r_n)\sigma_m^2 & (1 - r_n)\sigma_m^2 \end{pmatrix}. \quad \text{(A4)}$$

G has the same block-diagonal form as **P** but the first and third elements on the diagonal have σ_g^2 in place of σ_z^2 and r_n in place of t_n. Evaluation of Equations A1 and A2 using these formulas yields text Equations 8 and 9.

APPENDIX II

In a sample of N unrelated individuals from a population, consider the multiple regression of individual phenotype, z, on number of copies of a specific allele at each of a set of molecular marker loci significantly associated with the ith QTL, denoted as y_{ij} for the jth marker locus associated with the ith QTL,

$$z - \bar{z} = \sum_{j=1}^{k_i} \alpha_{ij}(y_{ij} - \bar{y}_{ij}) + \epsilon_i. \quad \text{(A5)}$$

The partial regression coefficient α_{ij} represents the additive effect on the specified allele at the jth marker locus associated with the ith QTL, and ϵ_i is an error term. When the k_i marker loci included in this regression have been determined *a priori* as those with the most highly significant partial regression coefficients from among all of the markers in the linkage group analyzed in the previous generation (e.g. by stepwise multiple regression), k_i will generally be a small number. The vector of partial regression coefficients then has the unbiased estimate $\hat{\alpha}_i = \mathbf{V}_i^{-1}\mathbf{C}_i$ in which $\mathbf{V}_i$ is the variance-covariance matrix of $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ik})^{\mathrm{T}}$, and $\mathbf{C}_i$ is the vector of covariances of $\mathbf{y}_i$ with z in the sample (KENDALL and STUART 1973 Ch. 19). In large samples the distribution of $\hat{\alpha}_i$ is asymptotically multivariate normal (SEARLE 1971 Ch. 3.5). The additive genetic variance at the ith QTL (or linkage group) associated with these marker loci, σ_i^2, is estimated without bias by

$$s_i^2 = \hat{\alpha}_i^{\mathrm{T}}\mathbf{V}_i\hat{\alpha}_i - k_i s_{\epsilon_i}^2/N \quad \text{(A6)}$$

where $s_{\epsilon_i}^2$ is an unbiased estimate of the error variance (KENDALL and STUART 1973 Ch. 28.12). The sampling variance of this expression is

$$\mathrm{Var}[s_i^2] = 4\sigma_i^2\sigma_{\epsilon_i}^2/N + 2k_i\sigma_{\epsilon_i}^4/N^2 + 2(k_i/N)^2\sigma_{\epsilon_i}^4/(N - k_i) \quad \text{(A7)}$$

in which $\sigma_{\epsilon_i}^2 = \sigma_z^2 - \sigma_i^2$ (SEARLE 1971 pp. 55–57, 99–100). From the assumption that $h^2p_i \ll 1$, we can set $\sigma_{\epsilon_i}^2 = \sigma_z^2(1 - h^2p_i) \cong \sigma_z^2$. The condition that the expected value of s_i^2 exceeds x_γ times its sampling variance then becomes approximately

$$1 > 2x_\gamma^2(Nh^2p_i)^{-1}\ \{2 + k_i[(N - k_i)h^2p_i]^{-1}\}. \quad \text{(A8)}$$

This produces an inequality for $(Nh^2p_i)^{-1}$ using the quadratic formula. Supposing that $4x_\gamma^2 \gg k_iN/(N - k_i)$, which generally will be the case when k_i is small (say up to 4 or 5) and N is large (since for example $x_{0.01} = 2.33$ and $x_{0.001} = 3.08$), yields the approximate condition (Equation 13a) in the text.

In a population sample grouped into ν families of size n, for a total of $N = \nu n$ individuals, the additive effects of alleles at k_i marker loci having highly significant associations with the ith QTL (or linkage group) can be estimated from separate multiple regressions of family mean phenotype, and of individual deviations from their family mean, on the numbers of specific alleles at the marker loci, analogous to Equation A5. This gives estimates $\hat{\alpha}_{if} = \mathbf{V}_{if}^{-1}\mathbf{C}_{if}$ and $\hat{\alpha}_{iw} = \mathbf{V}_{iw}^{-1}\mathbf{C}_{iw}$ corresponding to unbiased estimates of the additive genetic variance associated with the ith QTL between and within families of

$$s_{if}^2 = \hat{\alpha}_{if}^{\mathrm{T}}\mathbf{V}_{if}\hat{\alpha}_{if} - k_i s_{\epsilon_i f}^2/\nu$$

and (A9)

$$s_{iw}^2 = \hat{\alpha}_{iw}^{\mathrm{T}}\mathbf{V}_{iw}\hat{\alpha}_{iw} - k_i s_{\epsilon_i w}^2/(N - \nu).$$

In general the estimated additive genetic variance associated with the ith QTL, σ_i^2, is the sum of these components between and within families, but in a random mating population these components respectively are expected to equal $r_n\sigma_i^2$ and $(1 - r_n)\sigma_i^2$ (FALCONER 1981 Ch. 13). Because the estimates in Equation A9 are independent, they can be multiplied respectively by $1/r_n$ and $1/(1 - r_n)$ and combined additively, weighting the terms inversely by their sampling variances to achieve an unbiased estimate of σ_i^2 with minimal sampling variance (e.g. ROBERTSON 1955),

$$s_i^2 = \{c_f s_{if}^2/r_n + c_w s_{iw}^2/(1 - r_n)\}/(c_f + c_w) \quad \text{(A10)}$$

where

$$c_f = (1 - r_n)^{-2}\ \mathrm{Var}[s_{iw}^2] \quad \text{and} \quad c_w = r_n^{-2}\ \mathrm{Var}[s_{if}^2].$$

Since $\mathrm{Cov}[s_{if}^2, s_{iw}^2] = 0$, we have

$$\mathrm{Var}[s_i^2] = \mathrm{Var}[s_{if}^2]\mathrm{Var}[s_{iw}^2]/ \{(1 - r_n)^2\mathrm{Var}[s_{if}^2] + r_n^2\mathrm{Var}[s_{iw}^2]\}. \quad \text{(A11)}$$

Again assuming $h^2p_i \ll 1$, we can set $\sigma_{\epsilon_i f}^2 \cong t_n\sigma_z^2$ and $\sigma_{\epsilon_i w}^2 \cong (1 - t_n)\sigma_z^2$.

The general condition that the expected value of the weighted average s_i^2 exceeds x_γ times its sampling variance is somewhat cumbersome. To reduce it to a simple form, we assume that family sizes are large, $n - 1 \gg (1 - t)r_n/(1 - r)t_n$, so that almost all of the sampling variance is between families and Equation A10 becomes $s_i^2 \cong s_{iw}^2/(1 - r_n)$. The derivation then proceeds as before, with the additional assumption $4x_\gamma^2(1 - t)/(1 - r) \gg k_i(N - \nu)/(N - \nu - k_i)$ yielding the approximate condition (Equation 13b) in the text.

APPENDIX III

When the weighting coefficients in a selection index are obtained from estimates of the additive genetic and phenotypic variance-covariance matrices, denoted as $\hat{\mathbf{G}}$ and $\hat{\mathbf{P}}$, so that $\hat{\mathbf{b}} = \hat{\mathbf{P}}^{-1}\hat{\mathbf{G}}\mathbf{d}$, the realized gain in economic value in a large sample is

$$\mathbf{d}^{\mathrm{T}}\Delta\bar{\mathbf{z}} = \mathbf{d}^{\mathrm{T}}\mathbf{G}\hat{\mathbf{b}}i/\sqrt{\hat{\mathbf{b}}^{\mathrm{T}}\mathbf{P}\hat{\mathbf{b}}} \qquad \text{(A12)}$$

instead of Equation A2 (HARRIS 1964). Here the actual values of **G** and **P** are as in Equation A3 or A4. We assume that the standard quantitative genetic parameters are known exactly, and the only source of sampling error is in the total additive genetic variance associated with the marker loci, σ_m^2, which is estimated by s_m^2. The relative efficiency of MAS on individuals, analogous to Equation 4, can be expressed as a function of s_m^2 as

$$\xi(s_m^2) = \frac{\sigma_g^2 - s_m^2 + p\sigma_e^2}{\sqrt{(\sigma_g^2 - s_m^2)^2 + ph^2\sigma_e^2(\sigma_z^2 + \sigma_g^2 - 2s_m^2)}} \qquad \text{(A13)}$$

in which $\sigma_e^2 = \sigma_z^2 - \sigma_g^2$ is the environmental (plus nonadditive genetic) variance in the character. The expected relative efficiency in large samples can be approximated from a Taylor series as

$$\mathrm{E}[\xi(s_m^2)] = \xi(\sigma_m^2) + \tfrac{1}{2}\mathrm{Var}[s_m^2][\partial^2\xi(u)/\partial u^2]_{u=\sigma_m^2}. \qquad \text{(A14)}$$

Using the model of Equations 10–12 and Appendix II, we can define

$$s_m^2 = \sum_{i=1}^{l} s_i^2 \quad \text{and} \quad K = \sum_{i=1}^{l} k_i \qquad \text{(A15)}$$

and from (A7), assuming $k_i \ll N$, the sampling variance of s_m^2 is approximately

$$\mathrm{Var}[s_m^2] = 4\sigma_m^2\sigma_z^2/N + 2K\sigma_z^4/N^2. \qquad \text{(A16)}$$

Evaluating the second partial derivative of Equation A13 gives the expected proportional loss of efficiency, $1 - \mathrm{E}[\xi(s_m^2)]/\xi(\sigma_m^2)$, shown in text Equation 14.

A more rigorous derivation of (A13) considers the separate contributions of groups of marker loci associated with each of l QTLs detected. For markers associated with the ith QTL the molecular subscore of an individual, in the notation of Appendix II, is $m_i = \hat{\alpha}_i^{\mathrm{T}}\mathbf{y}_i$ and the unbiased estimate of the additive genetic variance associated with these markers is s_i^2. We wish to estimate the weight coefficients in the selection index

$$b_z z + b_1 m_1 + b_2 m_2 + \ldots + b_l m_l. \qquad \text{(A17)}$$

Assuming that the phenotypic and additive genetic variance in the character z are known exactly, and that the l QTLs detected are in linkage equilibrium with each other, we have

$$\hat{\mathbf{P}} = \begin{pmatrix} \sigma_z^2 & s_1^2 & s_2^2 & \ldots & s_l^2 \\ s_1^2 & s_1^2 & 0 & \ldots & 0 \\ s_2^2 & 0 & s_2^2 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ s_l^2 & 0 & 0 & \ldots & s_l^2 \end{pmatrix} \qquad \text{(A18)}$$

and $\hat{\mathbf{G}}$ has the same form but with σ_g^2 instead of σ_z^2 in the upper left corner. Evaluation of the estimated index weights in Equation A12 requires the matrix inverse

$$\hat{\mathbf{P}}^{-1} = \begin{pmatrix} \phi & -\phi & -\phi & \ldots & -\phi \\ -\phi & \phi + 1/s_1^2 & \phi & \ldots & \phi \\ -\phi & \phi & \phi + 1/s_2^2 & \ldots & \phi \\ \vdots & \vdots & \vdots & & \vdots \\ -\phi & \phi & \phi & \ldots & \phi + 1/s_l^2 \end{pmatrix} \qquad \text{(A19)}$$

in which $\phi = 1/(\sigma_z^2 - s_m^2)$ where s_m^2 is defined in Equation A15. This yields the estimated vector of index weights proportional to

$$\hat{\mathbf{b}}^{\mathrm{T}} = (\sigma_g^2 - s_m^2, \sigma_e^2, \sigma_e^2, \ldots, \sigma_e^2). \qquad \text{(A20)}$$

Evaluation of Equation A12 with economic values $\mathbf{d}^{\mathrm{T}} = (1, 0, 0, \ldots, 0)$ leads again to Equation A13.

Because each of the molecular subscores, m_i, has equal weighting in Equation A20, in the deterministic case when $s_m^2 = \sigma_m^2$ the selection index in Equation A17 reduces to the simple bivariate selection index in text Equation 2 with the molecular score m defined as the sum of the l molecular subscores for the separate QTLs.