Accuracy of marker-assisted selection with auxiliary traits

P NARAIN

29278 Glen Oaks Blvd. W., Farmington Hills, MI 48334-2932, USA

(Fax, 1-775-201-1282; Email, narainprem@hotmail.com)

Genetic information on molecular markers is increasingly being used in plant and animal improvement programmes particularly as indirect means to improve a metric trait by selection either on an individual basis or on the basis of an index incorporating such information. This paper examines the utility of an index of selection that not only combines phenotypic and molecular information on the trait under improvement but also combines similar information on one or more auxiliary traits. The accuracy of such a selection procedure has been theoretically studied for sufficiently large populations so that the effects of detected quantitative trait loci can be perfectly estimated. The theory is illustrated numerically by considering one auxiliary trait. It is shown that the use of an auxiliary trait improves the selection accuracy; and, hence, the relative efficiency of index selection compared to individual selection which is based on the same intensity of selection. This is particularly so for higher magnitudes of residual genetic correlation and environmental correlation having opposite signs, lower values of the proportion of genetic variation in the main trait associated with the markers, negligible proportion of genetic variation in the auxiliary trait associated with the markers, and lower values of the heritability of the main trait but higher values of the heritability of the auxiliary trait.

[Narain P 2003 Accuracy of marker-assisted selection with auxiliary traits; *J. Biosci*. **28** 569–579]

1. Introduction

The advent of the methods of molecular genetics has made significant impact on the statistico-genetic principles for the genetic improvement of domesticated plants and animals, as well as for the study of human diseases. Dispersed all along the chromosome, molecular markers such as those provided by restriction fragment length polymorphism (RFLP) have made it possible to detect and estimate the effects of quantitative trait loci (QTL) as shown by Lander and Botstein (1989) using maximum likelihood method. Genetic information of these markers has also been used as a criterion of indirect selection for genetic improvement of a given quantitative trait – a procedure of selection that has come to be known as marker-assisted selection (MAS). In fact, the value of blood group as a marker in predicting the breeding value of an individual was first realized by Neimann-Sorenson and Robertson (1961) in connection with the analysis of data on Danish cattle breeds. Smith (1967) showed how the selection on the basis of specific loci alone could be more efficient than the selection on the individual phenotype. The fact that such information can be integrated with those of artificial selection on individual and/or collateral basis in the form of an index to increase the efficiency of selection was demonstrated by the work of Lande and Thompson (1990) – herein after referred to as LT. The underlying basis of the MAS is the correlation between the trait and the marker genotype, which gets generated due to linkage disequilibria between the QTL and marker loci. Even in those cases where such loci are in linkage equilibrium and consequently no correlation between the trait and the marker exists at the population level, information on marker allele transmission could be included in the prediction of breeding values as shown by Fernando and Grossman (1989) using best linear unbiased prediction

Keywords. Accuracy; auxiliary traits; efficiency; index selection; marker-assisted selection; molecular correlations; molecular scores

Abbreviations used: MAS, Marker-assisted selection; QTL, quantitative trait loci; RFLP, restriction fragment length polymorphism.

(BLUP) methodology mostly applicable to animal breeding problems. Ollivier (1998) showed how information on marker allele transmission in a given family rather than on marker genotype could be used to predict the selection accuracy of MAS for populations in linkage equilibrium.

The increased efficiency of MAS, however, is accompanied by the increased cost involved in sample collection, DNA extraction and typing on the individuals in the sample as compared to that involved in taking simple measurements of the trait. The cost reduction for MAS can be achieved in several ways. New marker technologies such as those based on polymerase chain reaction (PCR) may reduce the cost of MAS. Selective genotyping of the extreme progeny, as advocated by Lander and Botstein (1989), is another way. When the cost of growing progeny is less than the cost of complete RFLP genotyping, it may be more efficient to increase the number of progeny grown but to genotype only those with the most extreme phenotypes because they provide the highest expected lod-score per progeny resulting in more linkage information. Yet another way could be to bring in auxiliary information from other traits that are correlated with the main trait and are cheaper to measure. This idea has been used in the past by several workers to increase the efficiency of individual and family selection itself by including in the index one or more auxiliary traits in conjunction with the main trait (Rendel 1954; Osborne 1957; Purser 1960; Searle 1965; Narain and Mishra 1975; Narain 1985; Narain and Kaur 1994). In particular, Narain and Mishra (1975) – hereinafter referred to as NM – showed how the efficiency of index selection can be increased when auxiliary information from related traits are included in the index.

As a matter of fact, molecular information in MAS is itself a sort of auxiliary information but obtained at a higher cost. It seems, therefore, that it may be worthwhile to examine how the accuracy of MAS due to index selection in LT compares with that in NM where the auxiliary information is usually obtained at a much lower cost. Also, given the index in NM, it may be of interest to examine how it's accuracy behaves if the information on the corresponding molecular scores, both for the main as well as auxiliary traits, is included in the index. In this paper, therefore, we propose such a new index for MAS with auxiliary traits and study its accuracy *vis-à-vis* those given in NM and LT.

2. Molecular scores

There is one aspect in which the molecular information is on a different footing than any measurable observed character. Genotyping a given individual by a molecular technique such as RFLP, gives the marker genotype; and regressing the observed values of a given trait on the marker genotype, treated as a dummy variable, gives the molecular score which can be treated as another character in the selection index. At the population level, this score, as a regression coefficient, is exactly known. It is, however, assumed that the QTL and the marker loci are in linkage disequilibrium.

We consider either a sufficiently large F_2, or a backcross population, following hybridization of selected lines for determining linkage relationships among a sufficient number of marker loci which are associated with linked QTL on the basis of linkage disequilibria between them for each of the linkage groups. As selection proceeds, these associations of marker loci are likely to change and therefore will require their re-evaluation every few or several generations. A multiple regression of individual phenotypic values carried out on the number of alleles at all the marker loci on a given chromosome will determine the additive genetic effects of QTLs associated with the linked markers. The predicted phenotypic value for a given individual obtained from this regression is then known as molecular score for that individual.

One can then choose those marker loci that have the largest significant additive genetic effects. In the subsequent generation of selection before the next hybridization, another multiple regression of the character values on the number of chosen marker loci can be performed. This gives unbiased estimates of additive genetic effects of QTLs which are associated with the chosen loci and which require to be scored only in each subsequent generation of selection until the need for re-evaluation of the association arises. As shown in LT, it is not necessary to include in the selection index separate contributions of groups of marker loci associated with each QTL detected. Instead, a net molecular score, defined as the sum of the molecular sub-scores for the separate QTLs, for each individual, need be included in the index.

3. Proposed selection index of MAS with auxiliary traits

We consider k auxiliary traits x_i, where, $i = 1, 2, \ldots, k$. These auxiliary traits are related to the main trait y with the corresponding phenotypic (P), genetic (g), net molecular (m) values associated with the chosen markers and the remainder (g') of the genetic values not associated with the chosen markers, all expressed as deviations from the population means. Then, if E denotes the corresponding environmental deviations, P is the sum of g and E and g is the sum of m and g'. Denoting the phenotypic and the molecular values for different characters, respectively, by $k \times 1$ column vector $\mathbf{P}(x)$, and scalar $P_0(y)$; $k \times 1$ column vector $\mathbf{m}(x)$ and scalar $m_0(y)$, the linear model for the selection index (I) proposed in this paper, is set-up as:

$$I = b_0\, P_0(y) + \mathbf{b}^T \mathbf{P}(x) + b_{0m}\, m_0(y) + \mathbf{b}_m^T\ \mathbf{m}(x), \qquad (1)$$

where $\mathbf{b}^T$ is the row vector of the coefficients b_1, b_2, ... b_k associated with the phenotypic values of the auxiliary traits, $\mathbf{b}_m^T$ is the row vector associated with the true molecular scores of the auxiliary traits, b_0 and b_{0m} are the respective coefficients for the phenotypic and molecular values of the main trait. These coefficients are to be so determined that the correlation between the index I and g_0 – the genetic value of the main trait – is maximum. In the genetic model of the index, this implies that: the phenotypic as well as the molecular values of the auxiliary traits, and the molecular values of the main trait, receive zero weights. On the other hand, the models for the indices in NM and LT respectively are:

$$I_{\mathrm{NM}} = b_{0(n)}\, P_0(y) + \mathbf{b}_{(n)}^T P(x); \text{ and} \qquad (2)$$

$$I_{\mathrm{LT}} = b_{0(t)}\, P_0(y) + b_{0m(t)}\, m_0(y). \qquad (3)$$

In (2) and (3), the b-coefficients have similar meanings as in (1), and n and t in parenthesis refer to the indices in NM and LT, respectively.

The derivation of the proposed selection index I together with related issues on the various correlations is given in the Appendix. The accuracy of selection based on index I – which is expressed as the maximized correlation coefficient between I and g_0 – and is denoted by $R_I^{(k)}$, is:

$$R_I^{(k)} = [h_0\,(1 - p_0^2 - b^T \Sigma^{*-1}\ a) + h_0\, Z(\, p_0^2/h_0^2 + a^T \Sigma^{*-1}\ a)] \cdot[(1 - p_0^2\ h_0^2 - b^T \Sigma^{*-1} b) + Z^2(p_0^2\ /h_0^2 + a^T \Sigma^{*\,-1}\ a)]^{-1/2}. \qquad (4)$$

The relative efficiency of I – which is denoted by $E_I^{(k)}$ – for a given intensity of selection compared to individual selection based on y alone with the same intensity of selection is then $R_I^{(k)}/h_0$. On the other hand, the accuracies of I_{NM} and I_{LT} are given by

$$R_{\mathrm{NM}}^{(k)} = [h_0\,(1 - \mathbf{r}_0^T\ \Sigma_P^{-1}\ \mathbf{c}(g)) + h_0\, Z_{\mathrm{N}}\, \mathbf{c}^T(g)\ \Sigma_P^{-1}\ \mathbf{c}(g)] \cdot[(1 - \mathbf{r}_0^T\ \Sigma_P^{-1}\ \mathbf{r}_0) + Z_{\mathrm{N}}^2\, \mathbf{c}^T(g)\ \Sigma_P^{-1}\ \mathbf{c}(g)]^{-1/2}. \qquad (5)$$

Here, $Z_{\mathrm{N}} = (1 - \mathbf{r}_0^T\ \Sigma_P^{-1}\ \mathbf{r}_0)/(1 - \mathbf{r}_0^T\ \Sigma_P^{-1}\ \mathbf{c}(g))$; and,

$$R_{\mathrm{LT}} = [p_0^2 + h_0^2\ f_0^4/(1 - p_0^2\ h_0^2)]^{1/2} = h_0[1 + p_0^2\ e_0^4/h_0^2(1 - p_0^2\ h_0^2)]^{1/2}. \qquad (6)$$

To compare the accuracies of I_{LT} and I, we express $R_I^{(k)}$ yet in another form as:

$$R_I^{(k)} = [R_{LT}^2 + h_0^2\ \{\mathbf{d}^T \Sigma^{*-1}\ \mathbf{d} + (1 - p_0^2\ h_0^2)\ b^T \mathbf{D} a\}/(1 - p_0^2\ h_0^2)\,(1 - p_0^2\ h_0^2 - b^T \Sigma^{*-1}\ b)]^{1/2}, \qquad (4a)$$

where,

$$\mathbf{d} = (1 - p_0^2\ h_0^2)\ a - f_0^2\, b = e_0^2\ f_0^2\ \{\mathbf{c}(g') - \mathbf{c}(\mathrm{e})\}; \text{ and}$$

$$\mathbf{D} = \Sigma^{*-1}\,(ab^T - ba^T)\Sigma^{*-1} = e_0^2\ f_0^2\ \Sigma^{*-1}\,\{\mathbf{c}(g')\mathbf{c}^T(e) - \mathbf{c}(e)\mathbf{c}^T(g')\}\ \Sigma^{*-1}.$$

The additional terms in the expression (4a) over and above the term involving the square of the accuracy of I_{LT} are clearly positive, indicating thereby that the gain is due to the use of auxiliary traits in MAS. However, the additional terms, depending upon a large number of parameters, could become zero when either $e_0^2 = 0$, or $f_0^2 = 0$, or else $\mathbf{c}(g') = \mathbf{c}(e)$ and hence in these three cases there would be no gain due to the use of auxiliary traits. In the first case, the main trait is totally heritable and the accuracy becomes one both for I as well as for I_{LT}. This indicates that no extra information can be provided either by the marker loci or by the auxiliary traits. In the second case, all the genetic variance is explained by the markers and there is no gain due to the use of auxiliary traits. In fact if we take p_0^2 very close to one, the maximum relative efficiency on the marker loci alone, compared to individual selection with the same intensity, is (p_0/h_0), the same as given in LT. The accuracy of selection is then just p_0. In the third case, the various parameters so adjust themselves that $\mathbf{c}(g')$ equals $\mathbf{c}(e)$ and the accuracy of the proposed index reduces to that of the I_{LT}.

If, however, we choose auxiliary traits which have no genetic correlations with the main trait and hence have no correlations at the molecular level, the accuracy becomes

$$[R_{LT}^2 + h_0^2\ f_0^4\ \mathbf{r}_0^T\ \Sigma^{*-1}\ \mathbf{r}_0/(1 - p_0^2\ h_0^2)\,(1 - p_0^2\ h_0^2 - \mathbf{r}_0^T\ \Sigma^{*-1}\ \mathbf{r}_0)]^{1/2}.$$

The additional term in this expression is again clearly positive, and therefore indicates the gain over the MAS with no auxiliary trait. In this form the use of the auxiliary traits in MAS amounts to correcting the main trait for their environmental effects, and the selection index reduces to the phenotypic index (Narain 1990).

4. MAS with one auxiliary trait

It is apparent from the theory presented in the previous section that a study of the behaviour of the efficiency or the accuracy of the MAS with variation in the various parameters for the k auxiliary traits is a formidable task unless we introduce some simplifications. We, therefore, assume that there is only a single auxiliary trait, i.e. $k = 1$. The vectors $\mathbf{r}_0$, $\mathbf{r}(m)$, $\mathbf{c}(g)$, $\mathbf{c}(e)$ and $\mathbf{c}(g')$ now become the scalars and can be represented as: r_{01}, $r_{01}(m)$, $c_1(g)$, $c_1(e)$ and $c_1(g')$ or simply as r_0, $r(m)$, $c(g)$, $c(e)$ and $c(g')$ respectively. The matrices Σ_P, Σ_R, $\mathbf{I}$, Π, $\mathbf{H}$, and $\mathbf{F}$ become the scalars which can now be represented as: 1, e_1^2, 1, p_1^2, h_1^2, and f_1^2 respectively. The vectors a, b, and the matrix Σ^* become respectively $[c(g) - c(m)p_0^2\ ] = f_0^2\ c(g')$, $(r_0 - c(m)\ p_0^2\ \ h_0^2\ ) = f_0^2\ \ h_0^2\ \ c(g') + e_0^2\ \ c(e)$, and $(1 - p_1^2\ \ h_1^2\ )$. Also, the matrix $\mathbf{D}$ becomes the scalar zero since ab^T

$= b a^T$, in view of the corresponding vectors becoming scalars. This gives, on simplification, for the accuracy of MAS with one auxiliary trait, $R_I^{(I)}$, denoted hereinafter simply by R_I, as

$$R_I = [R^2_{\mathrm{LT}} + h_0^2\, q^2/(1 - p_0^2\, h_0^2)\Delta]^{1/2}, \tag{7}$$

where,

$$q = e_0^2\, f_0^2\, \{c(g') - c(e)\}, \text{ and} \tag{8}$$

$$\Delta = (1 - p_0^2\, h_0^2)\,(1 - p_1^2\, h_1^2) - \{f_0^2\, h_0^2\, c(g') + e_0^2\, c\,(e)\}^2. \tag{9}$$

It is apparent from the expression (7) that if q is zero, there would be no gain due to the use of auxiliary trait x_1. This occurs when either $e_0^2 = 0$ i.e. $h_0^2 = 1$, or $f_0^2 = 0$ i.e. $p_0^2 = 1$, or else the difference $d = \{c(g') - c(e)\} = 0$. In the first case, the phenotypic correlation between y and x_1 is entirely due to the corresponding genetic correlation which is when $r_0 = r(g)\, h_1 = c(g)$. In that case R_{LT} would also be one, indicating that the individual selection on the phenotype would be as good as the genetic value itself and no extra information is provided either by the marker loci or by the auxiliary trait. In the second case, the entire additive genetic variance gets explained by the markers and the accuracies R_{LT} and R_{I} both again attain the maximum value of one. In the third case, the parameters $r(g')$, $r(e)$, h_0^2, h_1^2, p_0^2, and p_1^2 adjust themselves in a manner where the difference $d = 0$ giving the accuracy of the proposed index equal to R_{LT}. Of course this requires that the correlation coefficients $r(g')$ and $r(e)$ have the same signs; i.e. either both positive or both negative. But when they have opposite signs, the accuracy of the proposed index gets enhanced over R_{LT}. In fact as we will see numerically in the next section, the accuracy attains a minimum at the point when d is zero and increases on either of this point as d becomes positive or negative.

For the index I_{NM} the accuracy for $k = 1$, denoted by R_{NM}, reduces, from (5), to

$$R_{\mathrm{NM}} = h_0\,[1 + \{c(g) - r_0\}^2/(1 - r_0^2)]^{1/2}. \tag{10}$$

We can compare the expression (6) with that in (10) to see whether using auxiliary information in the form of molecular scores is better or not than that of using it as another correlated character. The two expressions will be equal if,

$$p_0^2 = [1/h_0^2\, \{1 + (1 - h_0^2)^2\,(1 - r_0^2)/h_0^4\, \{c(g) - r_0^2\}^2\}] = K, \text{ say,} \tag{11}$$

which is necessarily less than one but greater than zero. Therefore, for values of $p_0^2 < K$, the use of an auxiliary trait would be more efficient than the use of molecular markers, whereas for $p_0^2 > K$ the use of molecular information would be preferred. It may be noted that when $c(g) = r_0$, R_{NM} is just h_0 and $R_{\mathrm{LT}} > R_{\mathrm{NM}}$ for all the values of p_0^2.

We observe that even with $k = 1$, the efficiency is a function of as many as six parameters, viz. p_0^2, p_1^2, h_0^2, h_1^2, $r(e)$, and $r(g')$. We therefore introduce some simplification to reduce further the number of parameters. We consider the two traits to have heritabilities of the same order i.e. we assume $h_0^2 = h_1^2 = h^2$, denoting e_0^2 and e_1^2 by e^2. Then

$$R_I = [R^2_{\mathrm{LT}} + h^2\, q^{*2}/(1 - p_0^2\, h^2)\, \Delta^*]^{1/2}, \tag{12}$$

where

$$q^* = e^2\, f_0^2\,\{c(g') - c(\mathrm{e})\}, \tag{13}$$

$$\Delta^* = (1 - p_0^2\, h^2)\,(1 - p_1^2\, h^2) - \{\, f_0^2\, h^2\, c(g') + e^2\, c(e)\}^2. \tag{14}$$

If we choose an auxiliary trait which has very low proportion of additive genetic variance associated with the molecular markers i.e. if we take $p_1^2 = 0$, $f_1^2 = 1$, then $c(g')$ will become $r(g')/f_0$ and q^* and Δ^* will modify accordingly. The accuracy can then alternately be expressed as:

$$R_I = [R^2_{\mathrm{LT}} + \mathrm{h}^2\{(1 - p_0^2\, h^2)\, r(g) - (1 - p_0^2)r_0\}^2/ (1 - p_0^2\, h^2)\,(1 - p_0^2\, h^2 - r_0^2)]^{1/2}. \tag{15}$$

In this formula if we take the case of a main character whose entire genetic variance is associated with the markers i.e. $p_0^2 = 1$, the accuracy attains the maximum possible value of one, noting that $r(g)$, in that case, will be zero. However, if we put $p_0^2 = 0$ in this expression i.e. consider a trait for which the proportion of additive genetic variance associated with the markers is very low, noting that $r(g)$ is now not zero, we recover the expression (10) with $c(g)$ replaced by $r(g)$ as it should. On the other hand, if we choose an auxiliary trait whose entire genetic variance is associated with the markers, i.e. $p_1^2 = 1$, $f_1^2 = 0$, the accuracy would become as:

$$R_I = [R^2_{\mathrm{LT}} + h^2\, e^4\, f_0^4\, r^2\,(e)/(1 - p_0^2\, h^2)\{(1 - p_0^2\, h^2) (1 - h^2) - e^4\, r^2\,(e)\}]^{1/2}. \tag{16}$$

In this expression if we put $p_0^2 = 1$, $f_0^2 = 0$, then R^2_{LT} becomes one and the additional term zero so that the maximum possible accuracy of one is attained. However, if we take $p_0^2 = 0$, $f_0^2 = 1$, R^2_{LT} becomes h^2 and the accuracy reduces to $h[1 + r_0^2/(1 - h^2 - r_0^2)]^{1/2}$.

5. Numerical results with one auxiliary trait

The accuracies of MAS are evaluated for different combinations of the six parameters, i.e. p_0^2, p_1^2, h_0^2, h_1^2, $r(g')$ and $r(e)$, as given below using the expression (7).

p_0^2 : 0, 0·1, 0·2, 0·3, 0·4, 0·5, 0·6, 0·7, 0·8, 0·9, 1.

p_1^2 : 0, 0·5, 0·9, 1.

h_0^2 : 0·0015, 0·025, 0·05, 0·15, 0·45, 1.

h_1^2 : 0·0015, 0·025, 0·15, 0·45, 0·5, 1.

$r(g')$: – 1, – 0·9, – 0·8, – 0·7, – 0·6, – 0·5, – 0·4, – 0·3, – 0·2, – 0·1, 0, 0·1, 0·2, 0·3, 0·4, 0·5, 0·6, 0·7, 0·8, 0·9, 1.

$r(e)$: – 1, – 0·9, – 0·8, – 0·7, – 0·6, – 0·5, – 0·4, – 0·3, – 0·2, – 0·1, 0, 0·1, 0·2, 0·3, 0·4, 0·5, 0·6, 0·7, 0·8, 0·9, 1.

The results are presented in table 1 for different combinations of $r(g')$ and $r(e)$, when $p_0^2 = 0{\cdot}1$, $p_1^2 = 0{\cdot}5$, $h_0^2 = 0{\cdot}025$ and 0·45 (in bold face), and $h_1^2 = 0{\cdot}5$.

It is apparent from table 1 that first the accuracy is higher for the higher value (0·45) of h_0^2 as it should. Secondly, the accuracy 0·35 (**0·69**) for $[r(g'), r(e)] = (0, 0)$, corresponding to the index LT, are the lowest and those for the combinations (1, – 1) or (– 1, 1) are the highest and equal to the maximum possible of 1. The gain in the accuracy due to the use of the auxiliary trait is thus quite significant, particularly for the lower value of the heritability of the main trait. There is a rotational symmetry in the table, i.e. positive and negative combinations of the two correlations produce the same values. Furthermore, there are distinct values for 22 combinations (excluding the values of 1), out of 48 (excluding 0·0) given in the table. Thus as pointed out earlier, the accuracy is greater when correlations have opposite signs than when they have the same signs. It may be noted that if we want to see the effects of genetic correlation $r(g)$, molecular correlation $r(m)$ and phenotypic correlation r_0 on the accuracy for the given values of p_0^2, p_1^2, h_0^2, and h_1^2 then we have to convert the values of the particular combination $[r(g'), r(e)]$ to them using (A4) and (A5) for a given value of $r(m)$. For values of p_0^2 and p_1^2 close to zero, however, $r(g') = r(g)$, we need only (A4) to get the corresponding r_0, the parameter $r(m)$ getting eliminated in this case.

Table 1. Accuracy of MAS with one auxiliary trait for different combinations of $[r(g'), r(e)]$ when $p_0^2 = 0{\cdot}1$, $p_1^2 = 0{\cdot}5$, $h_0^2 = 0{\cdot}025/0{\cdot}45$, and $h_1^2 = 0{\cdot}5$.

	$r(g')$						
	– 1	– 0·8	– 0·5	0·0	0·5	0·8	1
$r(e)$							
– 1·0	1·00	0·74	0·45	0·40	0·69	0·88	1·00
	1·00	**0·77**	**0·79**	**0·85**	**0·92**	**0·96**	**1·00**
– 0·8	0·74	0·59	0·42	0·37	0·57	0·72	0·83
	0·69	**0·70**	**0·72**	**0·78**	**0·85**	**0·90**	**0·93**
– 0·5	0·65	0·54	0·42	0·35	0·49	0·62	0·71
	0·71	**0·70**	**0·69**	**0·72**	**0·78**	**0·82**	**0·86**
0·0	0·64	0·55	0·44	0·35	0·44	0·55	0·64
	0·77	**0·74**	**0·71**	**0·69**	**0·71**	**0·74**	**0·77**
0·5	0·71	0·62	0·49	0·35	0·42	0·54	0·65
	0·86	**0·82**	**0·78**	**0·72**	**0·69**	**0·70**	**0·71**
0·8	0·83	0·72	0·57	0·37	0·42	0·59	0·74
	0·93	**0·90**	**0·85**	**0·78**	**0·72**	**0·70**	**0·69**
1·0	1·00	0·88	0·69	0·40	0·45	0·74	1·00
	1·00	**0·96**	**0·92**	**0·85**	**0·79**	**0·77**	**1·00**

Values in bold face are for $h_0^2 = 0{\cdot}45$.

We now plot the accuracy as a function of the difference $d = c(g') - c(e)$ for the parameter values of $h_0^2 = 0{\cdot}025$, $h_1^2 = 0{\cdot}5$, $p_0^2 = 0{\cdot}1$, $p_1^2 = 0{\cdot}5$, and $r(e) = – 1, 0, 1$ (figure 1). As given in expression (7), the accuracy attains, the minimum at $d = 0$ as expected, all the three curves touch at this point and increased for positive or negative values of d on either side of this point. While the curve for $r(e) = 0$ is symmetrical around d, the curves for $r(e) = – 1$ and 1 are not symmetrical. The portions of the curve for $r(e) = – 1$ for positive d are the mirror images of those for $r(e) = 1$ for negative d. The values of accuracy for $r(e) = 0$ are all lower than those when $r(e) = – 1$ or 1. The common, as also the minimum, values of accuracy for $d = 0$ is 0·35 when $h_0^2 = 0{\cdot}025$ and 0·69 when $h_0^2 = 0{\cdot}45$ and correspond to the LT index with no auxiliary trait. Since all the other values of accuracy are greater than these values, it is apparent that the use of auxiliary trait improves the accuracy – provided the parameters do not have values which so adjust themselves as to make the value of d as zero. When the heritability of the main trait is 0·025, the accuracy of selection based on phenotypic values alone is only $\sqrt{0{\cdot}025} = 0{\cdot}158$ but increases to 0·35 when molecular information is incorporated via LT index which further increases to 0·72 with $r(g') = – 0{\cdot}8$ and $r(e) = 0{\cdot}8$ when an auxiliary trait is introduced in the index. It may however be noted that as $r(e)$ is fixed for a given curve, the variation of the accuracy with variation in d is actually a variation in the accuracy with the variation in $r(g')$.

We now plot the accuracy as a function of p_0^2 for values of $h_0^2 = 0{\cdot}0015$, 0·025, 0·15, 0·45 and 1, when $h_1^2 = 0{\cdot}025$, $p_1^2 = 0{\cdot}5$, $r(g') = – 0{\cdot}8$ and $r(e) = 0{\cdot}8$. The behaviour of the accuracy of I with variation in p_0^2 is shown in figure 2. It is apparent from this figure that the accuracy increases with the increase in p_0^2, the proportion of additive genetic variance associated with the markers for all values of h_0^2 except 1 when it has, as expected, the maximum possible value of 1 irrespective of the value of h_0^2. As given in expression (7), the accuracy attains, as expected, the minimum at $d = 0$ and all the three curves touch at this point, increasing for positive or negative values of d on either side of this point. The differences in the values of the accuracy, however, decrease as p_0^2 increases. The values of accuracy become zero when $p_0^2 = 1{\cdot}0$, which is the point of maximum accuracy. At the

starting point $p_0^2 = 0$, the accuracies are 0·201, 0·368, 0·635, 0·851, and 1 respectively for the five chosen values of h_0^2. This starting point corresponds to the case when no variation in the main trait is associated with the molecular markers. Since accuracy is h_0 times the efficiency, the maximum efficiency is 5·19. This corresponds to the lowest value of $h_0^2 = 0·0015$. In the LT index (no auxiliary trait), on the other hand, the efficiency is always 1 for all the values of h_0^2, when $p_0^2 = 0$. This initial gain in the efficiency due to the use of auxiliary trait, persists but with reduced magnitude as p_0^2 increases, and it vanishes at $p_0^2 = 1$ for all the chosen h_0^2 values. Thus the use of the auxiliary trait in the index over and above the inclusion of molecular information is found to be beneficial, particularly at lower values of p_0^2 and for lower values of h_0^2.

The behaviour of the accuracy with variation in p_0^2 for different values of h_1^2 as 0·0015, 0·025, 0·15, 0·45, and 1; when $h_0^2 = 0·025$, $p_1^2 = 0·5$, $r(g') = -0·8$ and $r(e) = 0·8$ (figure 3) shows similar pattern as that in figure 2 but with a difference. Herein, for the highest value of $h_1^2 = 1$, the accuracy is not constant at 1 but instead increases from 0·8 to 1 as p_0^2 increases from 0 to 1. The accuracy increases with increase in p_0^2 at other values of h_1^2 too. At the initial point on the axis of p_0^2 i.e. 0, the accuracies are 0·285, 0·368, 0·518, 0·674, and 0·802 respectively for the five selected values of h_1^2. This shows that for higher accuracy, a higher value of h_1^2 is needed. The maximum efficiency corresponding to the maximum accuracy of 1 is when $1/\sqrt{0·025} = 6·32$, and this is attained at $p_0^2 = 1$ for all the five curves.

Figure 4 shows the effect of p_1^2 on the variation in accuracy with variation in p_0^2 for $h_0^2 = 0·05$, $h_1^2 = 0·45$, $r(g') = -0·8$ and $r(e) = 0·8$. For all the four curves – corresponding to the four different values of $p_1^2 = 0$, 0·5, 0·9, and 1 – the accuracy increases with increase in p_0^2, and accuracy attain the maximum value of 1 at $p_0^2 = 1$ in all the cases. At a given value of p_0^2, the accuracy is however higher for a lower value of p_1^2. When $p_0^2 = 0$, the accuracies are respectively 0·753, 0·708, 0·573, and 0·357. A lower value of p_1^2 therefore gives rise to a higher accuracy. It is important to note that the accuracy value of 0.753 (corresponding to $p_1^2 = p_0^2 = 0$) is in fact the accuracy of I_{NM} for the given values of the other parameters. The figure reveals that for all the accuracy points on the various curves below this value, the index I_{NM} is more accurate than the proposed index I whereas on the contrary the index I is more accurate. It seems that the incorporation of molecular information in the index over and above the auxiliary trait would improve the accuracy only when the proportion of genetic variation in the auxiliary trait explained by the markers is negligible. This means the entire genetic correlation between the main and the auxiliary traits is not associated with the markers.

6. Discussion

It is clearly demonstrated in this paper that the use of auxiliary traits improves the accuracy of selection and hence the relative efficiency of index selection compared to individual selection for the same intensity of selection. It is particularly so for higher magnitudes of residual genetic correlation and environmental correlation having opposite signs, lower values of the proportions of genetic variation in the main trait associated with the markers,

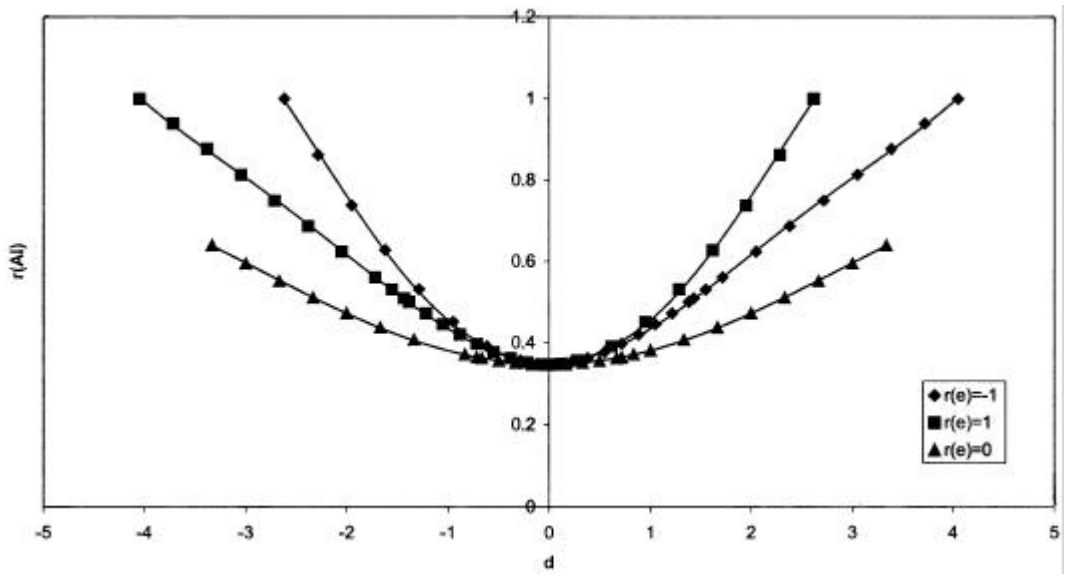

Figure 1. Accuracy of the index I with one auxiliary trait as a function of $d = c(g') - c(e)$ for $r(e) = -1$, 0, 1 when $p_0^2 = 0·1$, $p_1^2 = 0·5$, $h_0^2 = 0·025$, $h_1^2 = 0·5$.

negligible proportion of genetic variation in the auxiliary trait associated with the markers, and lower values of the heritability of the main trait but higher values of the heritability of the auxiliary trait. Additional recording of data on desirable auxiliary traits therefore can lead to cost reduction of MAS with increased efficiency.

While adopting this strategy there is however one issue which needs to be looked into. This is the sample size of the experiment considered here as very large so that the effects of detected QTLs are perfectly estimated. In practice this is not so (Beavis 1998; Utz *et al* 2000). As indicated in LT, the sampling errors in the parameter estimates reduce the efficiency of MAS by causing the weight coefficients to deviate from their optimal values. Moreau *et al* (1998) studied this in detail using simulations. They also showed that in samples of finite size the sampling errors which are associated with the weight coefficients and the effects which are associated with the markers, led to a smaller efficiency than expected under the assumption that the parameters are known. They pointed out the existence of an optimal heritability below which the low power of QTL detection and the bias caused by the selection of markers reduce the efficiency. In their model, samples of size N individuals are taken to be random samples from a large F_2 or a back-cross population following hybridization of selected lines for detecting marker-QTL associations on the basis of a simple regression of phenotypic values on marker types with a given probability of type I error. Each QTL is supposed to be linked to a single marker with observed rate of recombination between them assumed to be the same for all the marker-QTL pairs. The proportion of phenotypic variance associated with the additive effects of markers is estimated as the sum of adjusted R^2 at each marker that is significantly associated with a QTL with a given probability of type I error. This makes the parameter $p_0^2\ h_0^2$ (used in this paper) depend both on N and the number of QTLs considered. Also, the covariance between the genetic values of the QTL and the molecular score, based on the estimated additive effects associated with the markers, is not equal to the variance of the molecular score unless the additive effects of the QTLs are perfectly estimated without any sampling error. These factors affect the efficiency which gets reduced as shown. The efficiency is also shown to increase with N.

In a similar manner we can incorporate the effect of the finite sample size of the experiment when we have an auxiliary trait in addition to the main trait. If we have a main trait y and an auxiliary trait x, with corresponding QTL_y and QTL_x respectively, and both the traits are linked to the same marker m, we can then consider a triad m-QTL_y-QTL_x with observed recombination rates r_0 and r_1 between m-QTL_y and m-QTL_x respectively. r_0 and r_1 are assumed to be the same for each triad. The additive genetic effects can be estimated from simple regression of the trait value on the marker type separately for each of the two traits. The estimates over the traits would be correlated due to phenotypic and genetic correlation between the traits. The proportion of phenotypic variance for the ith trait associated with the additive effects of the

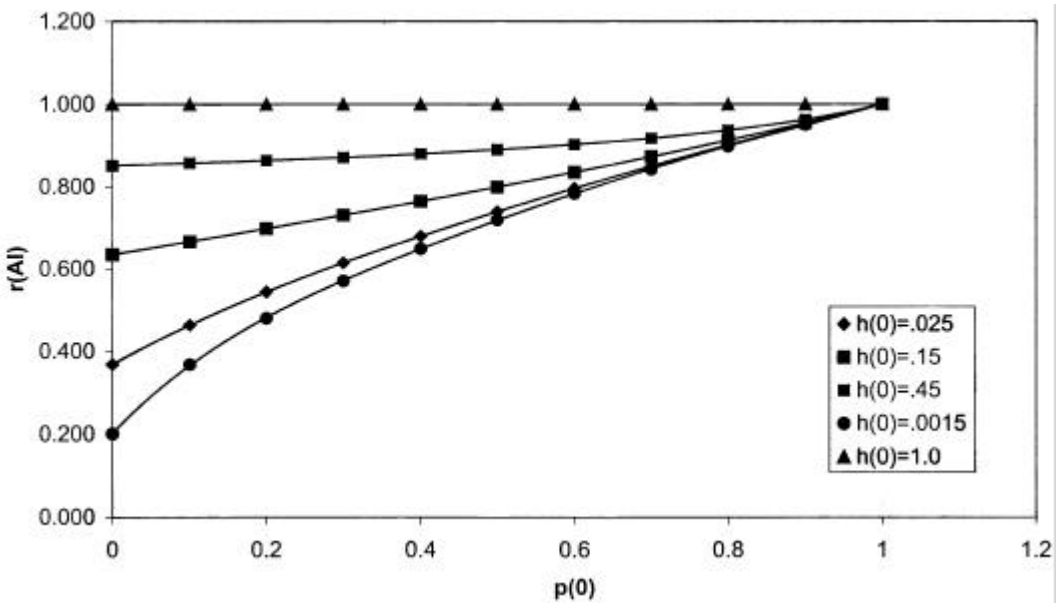

Figure 2. Accuracy of the index I with one auxiliary trait as a function of p_0^2 for h_0^2 = 0·0015, 0·025, 0·15, 0·45, 1 when p_1^2 = 0·5, h_1^2 = 0·025, $r(g')$ = – 0·8, $r(e)$ = 0·8.

markers can be estimated by the sum of the adjusted *R*-square, at each marker *m* significantly associated with a QTL with a given probability of type I error. This means, for $i = 0$ and 1, $p_0^2\ h_0^2$ and $p_1^2\ h_1^2$ are to be replaced respectively by

$$(N-1)/(N-2)\ \Sigma\ SS_m\ (i)/SS_{\text{tot}}\ (i) - k/(N-2),$$

where k is the number of QTL, SS_m denotes the sum of squares associated with the marker m, and SS_{tot} denotes the total sum of squares. For the covariance between the genetic value of the QTL and the molecular score for each of the two characters, we argue as below.

Let the molecular score for a given individual in the population for the *i*th trait be defined as

$$m\ (i) = \sum_q \hat{a}_q(i) q_q\ ,\ i = 0,\ 1, \tag{17}$$

where, $\hat{a}_q(i)$ is the estimated additive effects associated with the marker q for the *i*th trait and q_q is a dummy variable which takes the value 1 if the individual is of marker type MM, 0 if it is of type Mm and – 1 if it is of type *mm*. On the other hand if $d_k\ (i)$ denotes the additive effects of the *k*th QTL, the genetic values for the *i*th trait would be

$$g\ (i) = \sum_k d_k\ (i) q_k\ ,\ i = 0,\ 1. \tag{18}$$

The true additive effects associated with the marker q for the *i*th trait is:

$$a_q(i) = d_k(i)\mathrm{Cov}(q_k, q_q)/\mathrm{Var}(q_q)\ ,\ i = 0,\ 1. \tag{19}$$

Then the cross covariances between the traits:

$$\mathrm{Cov}[g(0),\ m(1)] = \sum_q a_q(0)\hat{a}_q(1)\mathrm{Var}(q_q),\ \text{and} \tag{20}$$

$$\mathrm{Cov}[g(1),\ m(0)] = \sum_q \hat{a}_q(0) a_q(1)\mathrm{Var}(q_q) \tag{21}$$

are obviously different. The cross covariances are also not equal to the covariance between the molecular scores for the two traits. The covariances between the molecular score for the two traits are given by:

$$\mathrm{Cov}[m(0),\ m(1)] = \sum_q \hat{a}_q(0)\hat{a}_q(1)\mathrm{Var}(q_q), \tag{22}$$

unless the additive effects associated with the marker are perfectly estimated with no sampling error i.e. $\hat{a}_q(i) = a_q(i)$ for $i = 0,\ 1$. Similarly the covariances within traits between g's and m's are not equal to the corresponding variances of m's due to the differences between the true and the estimated effects associated with the markers. In this paper we have assumed that these effects are truly known. How these modifications in the theory developed in this paper would affect quantitatively the accuracy and hence the efficiency of the MAS remains to be seen though it is apparent that there would be some loss of efficiency.

In the theory developed in this paper it is implicit that all the variation in the trait and the covariation between the traits explained by the molecular markers are genetic in origin. This need not necessarily be true (Whittaker 2001). The variation and covariation explained by the mar-

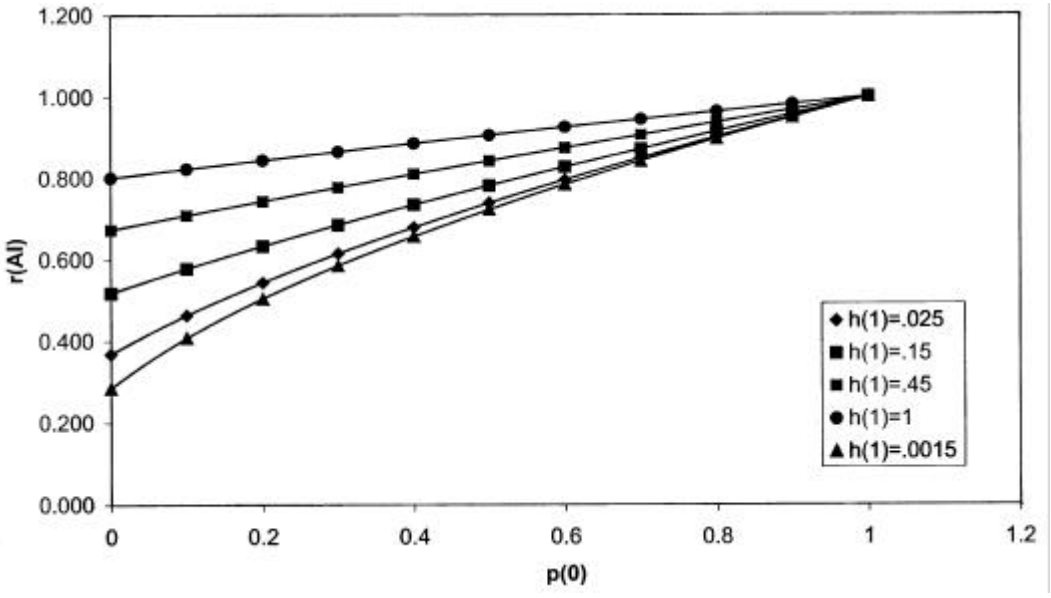

Figure 3. Accuracy of the index *I* with one auxiliary trait as a function of p_0^2 for h_1^2 = 0·0015, 0·025, 0·15, 0·45, 1 when p_1^2 = 0·5, h_0^2 = 0·025, $r(g')$ = – 0·8, $r(e)$ = 0·8.

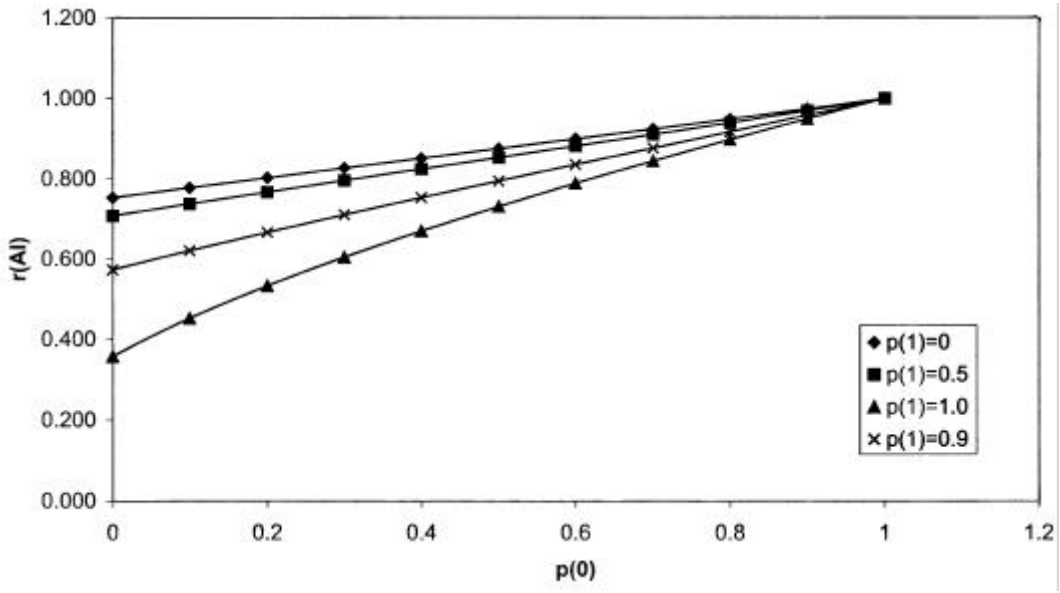

Figure 4. Accuracy of the index I with one auxiliary trait as a function of p_0^2 for p_1^2 = 0, 0·5, 0·9, 1 when h_0^2 = 0·05, h_1^2 = 0·45, $r(g')$ = – 0·8, $r(e)$ = 0·8.

kers as revealed by the multiple regression analysis may be partly genetic and the rest phenotypic. It seems this problem can only be handled by simulation studies of the type described in Whittaker (2001) and based on Whittaker *et al* (1997).

Acknowledgements

This study was undertaken when the author was Professor Emeritus at the Indian Agricultural Research Institute (IARI), New Delhi under an Emeritus Scientist Scheme entitled *Studies on Statistical Genetics and Evolution* of the Council of Scientific and Industrial Research (CSIR), New Delhi. He gratefully acknowledges the financial support received from the CSIR grants for undertaking this work. He is also grateful to the Director, IARI for providing the necessary facilities.

Appendix

Derivation of the accuracy of the proposed index I

Let the phenotypic, genetic, net molecular values associated with the markers and the remainder of the genetic values not associated with the markers, all expressed as deviations from the population means, be denoted by $P_i(x)$, $P_0(y)$, g_i, g_0, m_i, m_0, g_i' and g_0' respectively (the suffix i representing the ith auxiliary trait and the suffix 0 representing the main trait). The corresponding environmental deviations are denoted by E_i and E_0 respectively. Then, for each character, we have

$$P = g + E,$$
$$g = m + g'. \qquad \text{(A1)}$$

Also, let the phenotypic values be standardized to have unit variances so that the heritabilities of the auxiliary and main traits, h_i^2 and h_0^2 are the same as the respective genetic variances $s_g^2 s$. Let p_i^2 and p_0^2 denote the corresponding proportions of the genetic variances associated with the marker loci for the auxiliary and the main traits respectively. It may be noted that $p_i^2\ h_i^2$ and $p_0^2\ h_0^2$ would then be the corresponding proportions of the unit phenotypic variances associated with the marker loci for the auxiliary and the main traits respectively. Let r_{0i}, $r_{0i}(g)$, $r_{0i}(m)$ and $r_{0i}(e)$ denote respectively the phenotypic, genetic, molecular, and environmental correlations between the main trait and the auxiliary traits whereas $r_{0i}(g')$ be genetic correlation between the main trait and the ith auxiliary trait not associated with the markers for i = 1, 2, . . ., k. The phenotypic, genetic, molecular and environmental correlations between the auxiliary traits themselves are denoted respectively by r_{ij}, $r_{ij}(g)$, $r_{ij}(m)$, and $r_{ij}(e)$ whereas $r_{ij}(g')$ is the corresponding genetic correlation not associated with the markers for i, j = 1, 2, . . ., k. Then the covariances between different variables are given by:

$$\text{cov}(g_0, g_i) = r_{0i}(g)\ h_0 h_i;$$
$$\text{cov}(E_0, E_i) = r_{0i}(e)\ e_0\ e_i;$$

$$\text{cov}(m_0, m_i) = r_{0i}(m)\, h_0 h_i\, p_0 p_i;$$
$$\text{cov}(g'_0, g'_i) = r_{0i}(g')\, h_0\, h_i\, f_0\, f_i;$$
$$\text{cov}(g_i, g_j) = r_{ij}(g)\, h_i\, h_j;$$
$$\text{cov}(m_i, m_j) = r_{ij}(m)\, h_i\, h_j\, p_i\, p_j;$$
$$\text{cov}(g'_i, g'_j) = r_{ij}(g')\, h_i\, h_j\, f_i\, f_j; \text{ and}$$
$$\text{cov}(E_i, E_j) = r_{ij}(e)\, e_i\, e_j. \quad \text{(A2)}$$

Here,

$$e_0^2 = 1 - h_0^2,\ e_i^2 = 1 - h_i^2,\ f_0^2 = 1 - p_0^2,\ f_i^2 = 1 - p_i^2.$$

We assume that the covariances between g and E, between m and g', between m and E, as well as between g' and E are all zero for all the traits. We introduce the following parameters to express relationships among the various genetic, phenotypic, and environmental parameters:

$$c_i(g) = r_{0i}(g)\ (h_i/h_0);$$
$$c_i(e) = r_{0i}(e)\ (e_i/e_0);$$
$$c_i(m) = r_{0i}(m)\ (p_i h_i/p_0 h_0); \text{ and}$$
$$c_i(g') = r_{0i}(g')\ (f_i\, h_i/f_0\, h_0). \quad \text{(A3)}$$

It may be noted that because of the standardization and assuming the covariances cov (g, E), cov (m, g'), cov (m, E) and cov (g', E) to be zero, the phenotypic, genetic, and environmental correlations as well as the correlations associated and not associated with the molecular markers, between the main trait and the auxiliary traits for $i = 1, 2, \ldots k$ as well as between the auxiliary traits themselves for $i, j = 1, 2, \ldots k$ are mutually related by the following relations:

$$r_{0i} = c_i(g)\, h_0^2 + c_i(e)\, e_0^2; \quad \text{(A4)}$$
$$c_i(g) = c_i(m)\, p_0^2 + c_i(g')\, f_0^2; \quad \text{(A5)}$$
$$r_{ij} = r_{ij}(g)\, h_i\, h_j + r_{ij}(e)\, e_i\, e_j; \text{ and} \quad \text{(A6)}$$
$$r_{ij}(g) = r_{ij}(m)\, p_i\, p_j + r_{ij}(g')\, f_i\, f_j. \quad \text{(A7)}$$

It may be noted that when $p_0^2 = p_i^2 = p_j^2 = 1$, all of the additive genetic variances in the various traits are associated with the markers so that the genetic correlations between the traits are entirely determined by their corresponding molecular correlations giving $r_{0i}(m) = r_{0i}(g)$, $c_i(m) = c_i(g)$, $r_{ij}(m) = r_{ij}(g)$. But when p^2s are close to zero, the genetic correlations are mostly not associated with the markers so that $r_{0i}(g') = r_{0i}(g)$ and $c_i(g') = c_i(g)$. In either of the cases, we are left with only the relations (A4) and (A6) (Narain 1990).

Since the square of $r_{0i}(e)$ is less than unity, we have, from (A4), the condition, for each i as:

$$c_i(g)\, h_0^2 - e_i\, e_0 \le r_{0i} \le c_i(g)\, h_0^2 + e_i\, e_0. \quad \text{(A8)}$$

Similarly, since the square of $r_{0i}\ (g')$ is less than unity, we get, from (A5), the condition, for each i as:

$$c_i(m) p_0^2\, h_0^2 - f_i\, f_0\, h_i\, h_0 \le c_i(g)\, h_0^2 \le c_i(m) p_0^2\, h_0^2 + f_i\, f_0\, h_i\, h_0. \quad \text{(A9)}$$

Since $f_0^2\, h_0^2 = (1 - p_0^2)\, h_0^2 < (1 - p_0^2\, h_0^2)$, and $f_i^2\, h_i^2 = (1 - p_i^2)\, h_i^2 < (1 - p_i^2\, h_i^2)$, we have, for each i,

$$r_{0i} \le c_i(g)\, h_0^2 + e_i\, e_0,$$
$$< c_i(g)\, h_0^2,$$
$$< c_i(m)\, p_0^2\, h_0^2 + f_i\, f_0\, h_i\, h_0,$$
$$< c_i(m)\, p_0^2\, h_0^2 + \{(1 - p_0^2\, h_0^2)\,(1 - p_i^2\, h_i^2)\}^{1/2}.$$

Hence, both the conditions (A8) and (A9), taken together, give the condition, for each i,

$$c_i(m) p_0^2\, h_0^2 - \{(1 - p_0^2\, h_0^2)\,(1 - p_i^2\, h_i^2)\}^{1/2} < r_{0i} < c_i(m) p_0^2\, h_0^2 + \{(1 - p_0^2\, h_0^2)(1 - p_i^2\, h_i^2)\}^{1/2} \quad \text{(A10)}$$

which means, that for each i,

$$(1 - p_0^2\, h_0^2)\,(1 - p_i^2\, h_i^2) - (r_{0i} - c_i(m) p_0^2\, h_0^2)^2 > 0. \quad \text{(A11)}$$

It may, however, be noted that if we choose to express the quantities solely in terms of correlation coefficients $r_{0i}(e)$ and $r_{0i}(g')$ along with p's and h's, the various conditions stated above are not necessary, and the only obvious condition is that the values of these correlation coefficients should lie between + 1 and – 1.

Introducing vector and matrix notations, we let $\mathbf{r}_0$, $\mathbf{c}(g)$, $\mathbf{c}(e)$, $\mathbf{c}(m)$, and $\mathbf{c}(g')$ to denote respectively the $k \times 1$ vectors of r_{0i}, $c_i(g)$, $c_i(e)$, $c_i(m)$ and $c_i(g')$ and $\Sigma_{\mathbf{P}}$, Σ_g, Σ_E, Σ_m and $\Sigma_{g'}$ to denote respectively the $k \times k$ covariance matrices of r_{ij}(phenotypic), $r_{ij}(g)\, h_i h_j$(genetic), $r_{ij}(e)\, e_i\, e_j$ (environmental), $r_{ij}(m)\, p_i\, p_j\, h_i\, h_j$(molecular), and $r_{ij}(g')\, f_i\, f_j\, h_i\, h_j$ (genetic not associated with the markers). Then we let $\mathbf{H} = (\Sigma_g\, \Sigma_{\mathbf{P}}^{-1})$ and $\mathbf{E} = (\Sigma_E\, \Sigma_{\mathbf{P}}^{-1})$ (Narain 1990) and $\prod = (\Sigma_m\, \Sigma_g^{-1}) = (\Sigma_m\, \Sigma_{\mathbf{P}}^{-1}\, \mathbf{H}^{-1})$, $\mathbf{F} = (\Sigma_{g'}\, \Sigma_g^{-1}) = (\Sigma_{g'}\, \Sigma_{\mathbf{P}}^{-1}\, \mathbf{H}^{-1})$. The relations (A4), (A5), (A6), and (A7) now assume the following forms:

$$\mathbf{r}_0 = \mathbf{c}(g)\, h_0^2 + \mathbf{c}(e)\, e_0^2; \quad \text{(A12)}$$
$$\mathbf{c}(g) = \mathbf{c}(m)\, p_0^2 + \mathbf{c}(g')\, f_0^2; \quad \text{(A13)}$$
$$\Sigma_P = \Sigma_g + \Sigma_E; \text{ and} \quad \text{(A14)}$$
$$\Sigma_g = \Sigma_m + \Sigma_{g'}. \quad \text{(A15)}$$

We have then a condition, similar to (A11), i.e.

$$(1 - p_0^2\, h_0^2) - b^T\, \Sigma^{*-1}\, b > 0, \quad \text{(A16)}$$

where,

$$b = \mathbf{r}_0 - \mathbf{c}(m)\, p_0^2\, h_0^2 = h_0^2\, a + e_0^2\, \mathbf{c}(e) = f_0^2\, h_0^2\, \mathbf{c}(g') + e_0^2\, \mathbf{c}(e),$$
$$a = \mathbf{c}(g) - \mathbf{c}(m)\, p_0^2 = f_0^2\, \mathbf{c}(g'), \text{ and}$$
$$\Sigma^* = (\mathbf{I} - \prod \mathbf{H})\, \Sigma_{\mathbf{P}} = \Sigma_P - \Sigma_m = \Sigma_{g'} + \Sigma_E.$$

For the linear model for the selection index I set up in eq. (1) in the text, we maximize the correlation between I and g_0, i.e. the breeding value for the main trait to determine the coefficients optimally. This leads to the following set of normal equations:

$$b_0 + \mathbf{r}_0^T\,\mathbf{b} + p_0\,h_0^2\,b_{0m} + \mathbf{c}^T(m)\,\mathbf{b}_m\,p_0\,h_0^2 = h_0^2,$$
$$\mathbf{r}_0 b_0 + \boldsymbol{\Sigma}_\mathbf{P}\mathbf{b} + \mathbf{c}(m)\,p_0\,h_0^2\,b_{0m} + \prod\mathbf{H}\,\boldsymbol{\Sigma}_\mathbf{P}\,\mathbf{b}_\mathbf{m} = \mathbf{c}(g)h_0^2,$$
$$p_0\,h_0^2\,b_0 + \mathbf{c}^T(m)\,\mathbf{b}p_0\,h_0^2 + p_0\,h_0^2\,\,b_{0m} + \mathbf{c}^T(m)\,\mathbf{b}_m\,p_0\,h_0^2 = p_0\,h_0^2,\text{ and}$$
$$\mathbf{c}(m)\,p_0\,h_0^2\,b_0 + \boldsymbol{\Pi}\mathbf{H}\boldsymbol{\Sigma}_\mathbf{P}\,\mathbf{b} + \mathbf{c}\,(m)\,p_0\,h_0^2\,b_{0m} + \boldsymbol{\Pi}\mathbf{H}\boldsymbol{\Sigma}_\mathbf{P}\mathbf{b}_\mathbf{m} = \mathbf{c}(m)\,p_0\,\,h_0^2. \quad \text{(A17)}$$

It may be verified that if there were no auxiliary traits, these would reduce to

$$b_0 + p_0\,h_0^2\,\,b_{0m} = h_0^2\text{, and}$$
$$p_0\,h_0^2\,\,b_0 + p_0\,\,h_0^2\,\,b_{0m} = p_0\,\,h_0^2\,. \quad \text{(A18)}$$

This is the same set of equations given in LT for MAS. On the other hand, if there were no MAS, these would now reduce to

$$b_0 + \mathbf{r}_0^T\,\mathbf{b} = h_0^2\text{, and}$$
$$\mathbf{r}_0\,b_0 + \boldsymbol{\Sigma}_\mathbf{p}\,\mathbf{b} = \mathbf{c}\,(g)h_0^2. \quad \text{(A19)}$$

This is precisely the same set of equations as given in Narain (1990, p 396) for the auxiliary trait case.

Solving (A17), we get

$$b_0 = h_0^2/Z,$$
$$\mathbf{b} = h_0^2\,\,\boldsymbol{\Sigma}^{*-1}\,[a - b/Z],$$
$$b_{0m} = (1 - h_0^2\,/Z),\text{ and}$$
$$\mathbf{b}_\mathbf{m} = -\,h_0^2\,\boldsymbol{\Sigma}^{*-1}\,[a - b/Z], \quad \text{(A20)}$$

where,

$$Z = (1 - p_0^2\,h_0^2 - b^T\,\boldsymbol{\Sigma}^{*-1}\,b)/(1 - p_0^2 - b^T\boldsymbol{\Sigma}^{*-1}\,a). \quad \text{(A21)}$$

The accuracy of selection based on I, expressed as the maximized correlation coefficient between I and g_0, and denoted by $R_I^{(k)}$, is as represented in the text in eq. (4).

References

Beavis W D 1998 QTL analysis: power precision and accuracy; in *Molecular dissection of complex traits* (ed.) A H Paterson (Boca Raton, FL: CRC Press) pp 145–162

Fernando R L and Grossman M 1989 Marker assisted selection using best linear unbiased prediction; *Genet. Sel. Evol.* **21** 467–477

Lande R and Thompson R 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits; *Genetics* **124** 743–756

Lander E S and Botstein D 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps; *Genetics* **121** 185–199

Moreau L, Charcosset A, Hospital F and Gallais A 1998 Marker-assisted selection efficiency in populations of finite size; *Genetics* **148** 1353–1365

Narain P 1985 Progeny testing with auxiliary traits; *Biometrics* **41** 895–907

Narain P 1990 *Statistical genetics* (New York: John Wiley) (Wiley Eastern Ltd., New Delhi, reprinted in 1993. Published by the New Age International Pvt. Ltd., New Delhi in 1999)

Narain P and Mishra A K 1975 The efficiency of selective breeding based on a phenotypic index; *J. Genet.* **62** 69–76

Narain P and Kaur A P 1994 The accuracy of progeny testing with binomial auxiliary traits; *Anim. Prod.* **58** 189–196

Neimann-Sorensen A and Robertson A 1961 The association between blood groups and several production characteristics in three Danish cattle breeds; *Acta Agric. Scand.* **11** 163–196

Ollivier L 1998 The accuracy of marker-assisted selection for quantitative traits within populations in linkage equilibrium; *Genetics* **148** 1367–1372

Osborne R 1957 Correction for regression on a secondary trait as a method of increasing the efficiency of selective breeding; *Aust. J. Biol. Sci.* **10** 365–366

Purser A F 1960 The use of correction for regression on a second character to increase the efficiency of selection; in *Biometrical genetics* (ed.) O Kempthorne (London: Pergamon Press) pp 210–214

Rendel J M 1954 The use of regression to increase heritability; *Aust. J. Biol. Sci.* **7** 368–378

Searle S R 1965 The value of indirect selection. I Mass selection; *Biometrics* **21** 682–707

Smith C 1967 Improvement of metric traits through specific genetic loci; *Anim. Prod.* **9** 349–358

Utz H F, Melchinger A E and Schön C C 2000 Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples; *Genetics* **154** 1839–1849

Whittaker J C 2001 Marker-assisted selection and introgression; in *Handbook of statistical genetics* (eds) D J Balding, M Bishop and C Cannings (New York: John Wiley) pp 673–693

Whittaker J C, Haley C S, and Thompson R 1997 Optimal weighting of information in MAS; *Genet. Res.* **69** 137–144

MS received 16 December 2002; accepted 1 July 2003

Corresponding editor: JEAN R DAVID