Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 65(2) 2011 125-142

Multi-trait Multi-environment Genome-wide Molecular Marker Selection Indices

José Crossa[1*] and J. Jesús Cerón-Rojas[2]
[1]Biometrics and Statistics Unit of the Crop Research Informatics Laboratory (CRIL),
International Maize and Wheat Improvement Center (CIMMYT), México
[2]Colegio de Postgraduados, CP 56230, Km 36.5, Carretera México-Texcoco,
Montecillo, Edo. de México, México

Received 10 November 2010; Revised 25 February 2011; Accepted 28 February 2011

SUMMARY

Molecular selection indices such as the molecular eigen selection index method (MESIM) and other molecular marker score selection indices maximize the selection response by combining information on molecular markers linked to quantitative trait loci and phenotypic values of the traits. The standard restrictive selection index and the restrictive eigen selection index method (RESIM) maximize the selection response of only some traits while leaving others unchanged. This research extends the MESIM, RESIM, genomewide molecular selection index, and standard restrictive selection index to the case of a multitrait multienvironment genomewide molecular marker selection index. We used simulated data and real data for estimating the performance of various genomewide and molecular marker selection indices. Results showed that, in general, when several traits were selected in various environments simultaneously and all the markers were included in the indices, the multitrait multienvironment genomewide molecular marker selection index increased the genotypic means over the mean of individuals selected by other selection indices. The sampling properties of MESIM and RESIM in the context of multitrait multienvironment genomewide molecular marker selection indices and their selection responses are known, and their estimators showed desirable statistical properties such as consistency and asymptotic unbiasedness. We propose a general procedure for finding the asymptotic statistical sampling properties of the multitrait multienvironment genomewide molecular marker selection index and of other selection indices by applying the theory underlying MESIM and RESIM.

Keywords : Selection index, Restrictive selection indices, Eigenanalysis, Multitrait multienvironment genomewide molecular marker indices.

INTRODUCTION

In the context of marker-assisted selection, Lande and Thompson (1990) developed the theory of the molecular marker score selection index (LT), which is an application of the selection index methodology proposed by Smith (1936) but with the modification that the effects of the QTL linked to the markers (MQTL, fitted value of the least square regression) are incorporated into the index by means of a marker score (*i.e.*, MQTL × code values of the markers linked to QTL). In LT, the rate of genetic improvement is maximized by combining phenotypic information with the marker score. Gimelfarb and Lande (1994, 1995), and Zhang and Smith (1992, 1993) demonstrated that in large inbred populations, LT is very effective for traits with low heritability.

The efficiency of LT depends on the density of molecular markers (MM), the number of MM linked to QTL, the size of the population, and the heritability of the trait. Furthermore, since LT is derived from the

[*] *Corresponding author* : José Crossa
E-mail address : j.crossa@cgiar.org

selection index methodology proposed by Smith (1936), it has the same advantages and disadvantages as Smith's selection index: it is simple to use but its sampling statistical properties and selection response are unknown, except in the case of two traits (Hayes and Hill 1980). Even for two traits, the statistical properties of Smith's selection index and its selection responses are difficult to use and evaluate (Harris 1964); furthermore, it is not easy to consistently assign economic weights to the traits.

A different approach to the selection index is taken by Cerón-Rojas *et al*. (2006), who developed the basic theory of a selection index based on a dimension reduction method, the singular value decomposition (eigenanalysis) of the phenotypic variance-covariance (or correlation) matrix of the traits of interest (called ESIM, for eigen selection index method). The authors showed that ESIM does not require economic weights or estimates of genotypic variances-covariances. In ESIM the elements of the first eigenvector determine the proportion each trait contributes to the selection index, and the first eigenvalue is used in the selection response. Following the idea of the restrictive selection index of Kempthorne and Nordskog (1959), Cerón-Rojas *et al*. (2008a) developed a restrictive ESIM (RESIM) that facilitates maximizing the genetic progress of some characters while leaving others unchanged.

Recently, Cerón-Rojas *et al*. (2008b) developed a molecular marker score selection index (called MESIM, for molecular eigen selection index method), which, like the LT, maximizes the rate of genetic gains by combining the traits' phenotypic value with molecular scores. Like the ESIM, MESIM does not require economic weights but does require estimates of the genotypic variance-covariance structure of trait × trait, molecular score × molecular score, and trait × molecular score. In MESIM, the elements of the first eigenvector determine the proportion that each trait contributes to the selection index, and the first eigenvalue is used in the selection response. Simulation results of Cerón-Rojas *et al*. (2008b) showed that the genotypic means and the expected selection response from MESIM for each trait are equal to or greater than those from the LT. When several traits are simultaneously selected, MESIM performs well for traits with low heritability. An advantage of the MESIM over the LT is that the asymptotic statistical sampling properties of the estimators are well known.

In a typical plant breeding context where several traits are measured in a number of environments, MESIM can implicitly incorporate the phenomenon of genotype × environment interaction by including, as phenotypic variables, the combination of traits and environments (*i.e*., phenotypic variable = trait-environment combination) and the genetic correlations between environments, between traits, between traits and environments and all the associations between molecular marker scores, molecular scores and traits, and molecular scores and trait-environment combinations. Both LT and MESIM require identifying the linkage between the molecular marker and the QTL, the MQTL effects, and the combination of the molecular scores effects and phenotypic information that allows genotypes to be classified and selected using the selection index.

Linkage disequilibrium between MM and QTLs can be used to improve the prediction of individual performance and/or to map QTLs. There is evidence that not all QTLs that control traits may be effectively detected (Li 1998); therefore, a possible solution to improve the ability to detect chromosome regions that significantly contribute to the traits' phenotypic variability is provided by a genomewide molecular marker selection index that does not require first identifying a subset of markers with significant effects (Lange and Whittaker 2001; Bernardo and Yu 2007). In genomewide selection, traits and MM can be considered variables of the selection index that are used to predict the overall genetic merits of the individuals subject to selection. In this situation it is not necessary to estimate the MQTL effects, and as Lange and Whittaker (2001) pointed out, a genomewide marker selection index using all available markers is superior to the Lt. Bernardo and Yu (2007) found that, in maize, genomewide marker selection yielded genetic gains 18-43% above those achieved using LT.

In the context of molecular marker genomewide selection, a huge number of collinear MM are used as variables in the hope that they (or at least some of them) will, to some degree, be able to explain the observed phenotypic variability, thus increasing their ability to predict genotypic response. However, various statistical problems hindered an appropriate assessment of their importance in explaining the phenotypic response. Gianola *et al*. (2003, 2006) and Gianola and van Kaam (2008) pointed out statistical and genetic difficulties encountered in the process of building a model for genotypic prediction using the ordinary least square estimation method. Among these drawbacks are: (1) difficulty in handling non-additivity (*i.e.*, dominance, epistatic effects) in a standard parametric model even with two loci; (2) problems of false positives when multiple testing; (3) rigid assumptions, *i.e.*, linear relationships between response and explanatory variables; (4) collinearity of the MM and, therefore, no precise estimates of their effects; and (5) a number of MM far exceeding the number of observations. Gianola and van Kaam (2008) developed a theoretical background for integrating some non-parametric and semiparametric models that allow accounting for non-additivity effects (without explicitly modeling them) with Fisher's traditional infinitesimal additive genetic model.

One of the main problems in building models when using large numbers of linked markers is parameter identification. In the context of association mapping methodology, Crossa *et al*. (2007) pointed out aliasing problems due to ambiguities in the statistical model that are often encountered when including large numbers of markers, or when the researcher attempts to model the effect of genotype × environment interaction together with the genetic covariance among relatives and population structure. The problems of parameter identification created by large, collinear numbers of MM can be, at least partially, mitigated by the dimension reduction methodology of singular value decomposition (*i.e*., eigenanalysis). Since MESIM and RESIM are based on subjecting the high order dimension variance-covariance matrix (or correlation) comprising variable × variable, molecular score × molecular score, and variable × molecular score to singular value decomposition, thus extracting the main part of the signal in the early components, it seems that MESIM and RESIM might be an efficient mathematical approach to address the complexity of predicting genetic values via incorporating: (1) extensive genomewide MM information, and (2) large phenotypic data sets comprising multitraits measured in multienvironments.

In this paper, we extend the theory underlying the MESIM of Cerón-Rojas *et al*. (2008b) and the Lange and Whittaker (2001) selection index to the case of multitrait multienvironment genomewide molecular marker selection index. Furthermore, we show that this theory is also valid in the context of the restrictive selection index of Kempthorne and Nordskog (1959), and for the RESIM of Cerón-Rojas *et al*. (2008a).

The application of MESIM and the Lange and Whittaker (2001) selection index to the case of multitrait multienvironment molecular marker genomewide will be named MESIM$_{GW}$ (GW for genome wide) and LW (for Lange and Whittaker), respectively, whereas the application of the restrictive selection index of Kempthorne and Nordskog (1959) and the RESIM of Cerón-Rojas *et al*. (2008a) to the case of multitrait multienvironment molecular marker genomewide will be named KN$_{GW}$, and RESIM$_{GW}$, respectively. In all cases, the phenotypic data include multitrait multienvironment, and the genotypic data include all available molecular markers, not just those linked to QTLs, for example, when applying MESIM and LT. For MESIM$_{GW}$, LW, KN$_{GW}$, and RESIM$_{GW}$, the code values of the homozygous genotypes, markers, and QTLs are denoted by –1, and 1; furthermore, no scores are given to the markers, since in most molecular marker genomewide instances, the chromosomal locations of the markers and QTLs are unknown. We use (1) simulated data of doubled haploids for estimating the performance of MESIM$_{GW}$, LW, KN$_{GW}$, RESIM$_{GW}$, MESIM, and LT when phenotypic data and MM are used, and (2) real data from a QTL mapping study in an F$_3$ maize population. We propose a general procedure for finding the asymptotic statistical sampling properties of LW, KN$_{GW}$, and LT by applying the theory underlying MESIM$_{GW}$, RESIM$_{GW}$, and MESIM.

THEORY OF SELECTION INDICES

Although the theories underlying various selection indices are given in several publications (Lande and Thompson 1990, Falconer and Mackay 1997, Lange and Whittaker 2001, Bernardo 2002), here we will briefly describe and summarize the selection indices to facilitate an understanding of how they were extended to the case of multitrait multienvironment genomewide marker selection indices.

Smith's selection index : Smith's selection index is based on the linear combinations

$$Y = \boldsymbol{\beta}'\mathbf{p} \text{ and } Z = \boldsymbol{\theta}'\mathbf{g} \qquad (1)$$

where **p** is the vector of trait phenotypic values, **g** is the vector of trait genotypic values, $\boldsymbol{\beta}$ is the vector of the coefficients of $\mathbf{p}$, Z is the breeding value, and θ is the vector of trait economic weights. The trait phenotypic values, p_{it}(t = 1, 2,…, N_t, N_t = number of traits under selection, i = 1, 2,…, n, n = number of genotypes or sample size) are modeled as $p_{it} = g_{it} + \varepsilon_{it}$, where g_{it} is the i^{th} genotypic value of the t^{th} trait phenotypic value and ε_{it} is the environmental component. Assuming that g_{it} and ε_{it} are independent, and that g_{it} represents only additive effects, $Z = \theta'\mathbf{g}$ denotes the breeding value (Hazel 1943; Kempthorne and Nordskog 1959). Hence, selection based on $Y = \boldsymbol{\beta}'\mathbf{p}$ leads to a selection response

$$R = k\sigma_Z \rho_{YZ} = k\sigma_Z \frac{\boldsymbol{\theta}'\mathbf{G}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\theta}'\mathbf{G}\boldsymbol{\theta}}\sqrt{\boldsymbol{\beta}'\mathbf{P}\boldsymbol{\beta}}} \tag{2}$$

where $\mathbf{G}$ and $\mathbf{P}$ are the variance-covariance matrices of genotypic and phenotypic values, respectively, k is the standardized selection differential, $\theta'\mathbf{G}\beta$ is the covariance between Y and Z, $\boldsymbol{\beta}'\mathbf{P}\boldsymbol{\beta}$ is the variance of Y, $\sigma_Z^2 = \boldsymbol{\theta}'\mathbf{G}\boldsymbol{\theta}$ is the variance of Z, and ρ_{YZ} is the correlation between Y and Z.

The restrictive selection index of the Kempthorne and Nordskog method (KN RSI): Kempthorne and Nordskog (1959) maximize ρ_{YZ}^2 and incorporate restrictions into the genotypic variance-covariance matrix. Suppose there are q traits and only $q - r$ are to be improved while leaving r of them unchanged. Then, the index $Y = \boldsymbol{\beta}'\mathbf{p}$ should maximize ρ_{YZ}^2 while leaving r traits unchanged. Suppose a $q \times r$ matrix $\mathbf{F}$ of 0s and 1s such that $\boldsymbol{\beta}'\mathbf{GF} = \mathbf{0}$ [where the 1s are used for traits that remain unchanged, or fixed]. Let $\mathbf{C} = \mathbf{GF}$; according to Bulmer (1980), maximizing ρ_{YZ}^2 is the same that maximizing $\theta'\mathbf{G}\boldsymbol{\beta}$ under the restrictions $\boldsymbol{\beta}'\mathbf{P}\boldsymbol{\beta} = 1$ and $\boldsymbol{\beta}'\mathbf{C} = \mathbf{0}$. Then it is necessary to maximize $\Psi = \theta'\mathbf{G}\boldsymbol{\beta} - 0.5\tau(\boldsymbol{\beta}'\mathbf{P}\boldsymbol{\beta} - 1) - \mathbf{u}'\mathbf{C}'\boldsymbol{\beta}$, where 0.5τ and $\mathbf{u}' = [u_1 \ldots u_r]$ are Lagrange multipliers. When partial derivatives of Ψ with respect to $\boldsymbol{\beta}$ are set equal to the null vector, then

$$\mathbf{G}\boldsymbol{\theta} - \tau\mathbf{P}\boldsymbol{\beta} - \mathbf{Cu} = \mathbf{0}$$

Thus, the vector of KN RSI coefficients that maximizes $\theta'\mathbf{G}\boldsymbol{\beta}$ (and thus ρ_{YZ} and R) is $\boldsymbol{\beta}_{KN} = \mathbf{A}\boldsymbol{\beta}_S$ (the subscript KN in $\boldsymbol{\beta}_{KN}$ stands for Kempthorne and Nordskog), where $\mathbf{A} = [\mathbf{I} - \mathbf{P}^{-1}\mathbf{C}(\mathbf{C}'\mathbf{P}^{-1}\mathbf{C})^{-1}\mathbf{C}']$, $\mathbf{P}^{-1}$ is the inverse matrix of $\mathbf{P}$, and $\boldsymbol{\beta}_S = \mathbf{P}^{-1}\mathbf{G}\theta$. Thus, in the context of the KN RSI, the Y_{KN} that maximizes ρ_{YZ} is $Y_{KN} = \boldsymbol{\beta}'_{KN}\mathbf{p}$, and the maximized selection response is $R_{KN} = k\sqrt{\boldsymbol{\beta}'_{KN}\mathbf{P}\boldsymbol{\beta}_{KN}}$.

The restrictive eigen selection index method (RESIM): Similar to KN RSI, RESIM (Cerón-Rojas *et al.* 2008a) maximizes ρ_{YZ}^2, but besides the usual restrictions of KN RSI (*i.e.*, $\boldsymbol{\beta}'\mathbf{P}\boldsymbol{\beta} = 1$ and $\boldsymbol{\beta}'\mathbf{C} = \mathbf{0}$), it incorporates the restriction $\theta'\mathbf{G}\theta = 1$. Then, ρ_{YZ}^2 must be maximized under three restrictions: $\boldsymbol{\beta}'\mathbf{P}\boldsymbol{\beta} = 1$, $\theta'\mathbf{G}\theta = 1$, and $\boldsymbol{\beta}'\mathbf{C} = \mathbf{0}$. Therefore, it is necessary to maximize

$$\Theta = (\theta'\mathbf{G}\boldsymbol{\beta})^2 - \mathbf{u}'\mathbf{C}'\boldsymbol{\beta} - \mu(\boldsymbol{\beta}'\mathbf{P}\boldsymbol{\beta} - 1) - \omega(\theta'\mathbf{G}\theta - 1)$$

with respect to $\boldsymbol{\beta}$, θ, $\mathbf{u}' = [u_1 \ldots u_r]$, μ, and ω, where $\boldsymbol{\beta}$ is the vector of RESIM coefficients, θ is the vector of economic weights, and $\mathbf{u}' = [u_1 \ldots u_r]$, μ, and ω are Lagrange multipliers. The solution is

$$(\mathbf{Q}_R - \omega\mathbf{I})\,\boldsymbol{\beta} = \mathbf{0}$$

where $\mathbf{Q}_R = [\mathbf{I} - \mathbf{P}^{-1}\mathbf{C}(\mathbf{C}'\mathbf{P}^{-1}\mathbf{C})^{-1}\mathbf{C}']\mathbf{P}^{-1}\mathbf{G}$. Thus, the values that maximize ρ_{YZ}^2 under the three restrictions $\boldsymbol{\beta}'\mathbf{P}\boldsymbol{\beta} = 1$, $\theta'\mathbf{G}\theta = 1$, and $\boldsymbol{\beta}'\mathbf{C} = \mathbf{0}$ is the first eigenvalue (ω), and the vector that allows constructing $Y = \boldsymbol{\beta}'\mathbf{p}$ in RESIM is the first eigenvector ($\boldsymbol{\beta} = \boldsymbol{\beta}_{RESIM}$) of the matrix $\mathbf{Q}_R$, *i.e.*, $Y = \boldsymbol{\beta}'_{RESIM}\,\mathbf{p}$.

Lande and Thompson molecular selection index (LT): Lande and Thompson (1990) extended Eq. 1 to include the case where information on QTLs associated with molecular markers is available. They denoted the molecular score selection index as

$$Y_z = \boldsymbol{\beta}'_p\mathbf{p} + \boldsymbol{\beta}'_s\mathbf{s} = [\boldsymbol{\beta}'_p \quad \boldsymbol{\beta}'_s]\begin{bmatrix}\mathbf{p}\\ \mathbf{s}\end{bmatrix} \tag{3}$$

where $\boldsymbol{\beta}_p$ is a vector of trait phenotypic weights, $\boldsymbol{\beta}_s$ is the vector of the molecular score weights, $\mathbf{p}$ is the vector of trait phenotypic values, and $\mathbf{s} = [s_1 \ldots s_{N_s}]$, where each s_l (l =1, 2, …, N_s; N_s = number of molecular scores) is the l^{th} molecular score given by the sum of the products of the estimated MQTL effects multiplied by the coded values of their corresponding molecular markers. The selection response to the LT is

$$R = k\sigma_s\rho_{Y_sZ_s} = k\sigma_s \frac{\boldsymbol{\theta}'_{gs}\mathbf{W}\boldsymbol{\beta}_{ps}}{\sqrt{\boldsymbol{\theta}'_{gs}\mathbf{W}\boldsymbol{\theta}_{gs}}\sqrt{\boldsymbol{\beta}'_{ps}\mathbf{T}\boldsymbol{\beta}_{ps}}} \tag{4}$$

where $\mathbf{T} = \begin{bmatrix}\mathbf{P} & \mathbf{S}\\ \mathbf{S} & \mathbf{S}\end{bmatrix}$ and $\mathbf{W} = \begin{bmatrix}\mathbf{G} & \mathbf{S}\\ \mathbf{S} & \mathbf{S}\end{bmatrix}$; k has been defined as in Eq. 2; $\sigma_s^2 = \boldsymbol{\theta}'_{gs}\mathbf{W}\boldsymbol{\theta}_{gs}$ is the variance of the breeding value ($Z_s = \boldsymbol{\theta}'_g\mathbf{g} + \boldsymbol{\theta}'_s\mathbf{s}$); $\boldsymbol{\theta}'_{gs} = [\boldsymbol{\theta}'_g \quad \boldsymbol{\theta}'_s]$ is a vector of economic weights (in the LT selection index, θ_s is a vector of zeros); $\boldsymbol{\beta}'_{ps}\mathbf{T}\boldsymbol{\beta}_{ps}$ is the variance of Y_s;

$\boldsymbol{\beta}'_{ps} = [\boldsymbol{\beta}'_p \quad \boldsymbol{\beta}'_s]$ is a vector containing trait phenotypic (β_p) and molecular (β_s) weight scores; **G** and **P** are the variance-covariance matrices defined in Eq. 2; and **S** = Var(**s**) is the variance-covariance matrix of the molecular marker scores when two or more traits are considered. Only statistically significant additive MQTL effects are included in **s**.

The vector $\beta_{ps} = \mathbf{T}^{-1}\mathbf{W}\theta_{gs}$ allows constructing the molecular score LT selection index $Y_s = \boldsymbol{\beta}'_{ps}\mathbf{p}_{ps}$, $\mathbf{p}'_{ps} = [\mathbf{p}'\,\mathbf{s}']$, which has maximum correlation with $Z_s = \boldsymbol{\theta}'_g\mathbf{g} + \boldsymbol{\theta}'_s\mathbf{s}$; the maximized selection response is $R_s = k\sqrt{\boldsymbol{\beta}'_{sp}\mathbf{T}\boldsymbol{\beta}_{sp}}$. Estimators of β_p and β_s ($\hat{\boldsymbol{\beta}}_p$ and $\hat{\boldsymbol{\beta}}_s$) for various traits are obtained directly from the estimators of **G**, **P**, and **S** ($\hat{\mathbf{G}}$, $\hat{\mathbf{P}}$, and $\hat{\mathbf{S}}$), and from the vector θ_{gs}.

The molecular eigen selection index method (MESIM): Using a concept similar to that of Kempthorne and Nordskog (1959), Cerón-Rojas *et al.* (2008b) showed that Eq. 4 is maximized by maximizing $\rho^2_{Y_z Z_s}$. In MESIM, it is necessary to maximize

$$\Phi = (\boldsymbol{\theta}'_{gs}\mathbf{W}\boldsymbol{\beta}_{ps})^2 - \mu(\boldsymbol{\beta}'_{ps}\mathbf{T}\boldsymbol{\beta}_{ps} - 1) - \omega(\boldsymbol{\theta}'_{gs}\mathbf{W}\boldsymbol{\theta}_{gs} - 1)$$

with respect to β_{ps}, θ_{gs}, μ, and ω, where β_{ps} is the vector of MESIM coefficients, θ_{gs} is the vector of economic weights, μ and ω are Lagrange multipliers, and $\beta'_{ps}\mathbf{T}\beta_{ps} = 1$ and $\theta'_{gs}\mathbf{W}\theta_{gs} = 1$ are restrictions impose when maximizing $\rho^2_{Y_s Z_z}$. The result is

$$(\mathbf{K} - \mu\mathbf{I})\beta_{ps} = 0 \tag{5}$$

where $\mathbf{K} = \mathbf{T}^{-1}\mathbf{W}$. Thus, for MESIM, the value that maximizes $\rho^2_{Y_s Z_s}$ is the first eigenvalue (μ) of matrix **K**, and the vector that allows constructing Y_s (with maximum correlation to $Z_s = \boldsymbol{\theta}'_g\mathbf{g} + \boldsymbol{\theta}'_s\mathbf{s}$) is the first eigenvector of matrix **K** ($\beta_{ps} = \beta_{MESIM}$); the maximized selection response can be written as $R_{\text{MESIM}} = k\sqrt{\mu}$. When MQTL effects are not incorporated into the selection index, Eq. 5 can be written as $(\mathbf{P}^{-1}\mathbf{G} - \mu\mathbf{I})\beta = \mathbf{0}$ and then $\mathbf{K} = \mathbf{P}^{-1}\mathbf{G}$ (**G** and **P** were defined in Eq. 2).

In MESIM the estimates of μ and $\beta_{ps} = \beta_{MESIM}$ are obtained by singular value decomposition, which means **K** can be written as

$$\mathbf{K} = \mathbf{L}\boldsymbol{\Delta}\mathbf{H}' \tag{6}$$

where the columns of matrix $\mathbf{L}(\mathbf{L}\mathbf{L}' = \mathbf{I})$ are the left singular vector of **K**, and the columns of matrix $\mathbf{H}(\mathbf{H}\mathbf{H}' = \mathbf{I})$ are the right singular vector of **K;** Δ is a diagonal matrix with the square root of the eigenvalues (singular values) of $\mathbf{K}\mathbf{K}'$ or $\mathbf{K}'\mathbf{K}$. Then the estimators of $\mu = \mu_{MESIM}$ and $\beta_{ps} = \beta_{MESIM}$ are obtained from $\hat{\mathbf{K}}\hat{\mathbf{K}}'$, such that $(\hat{\mathbf{K}}\hat{\mathbf{K}}' - \hat{\mu}^2_{MESIM}\mathbf{I})\hat{\boldsymbol{\beta}}_{MESIM} = \mathbf{0}$; $\hat{\mu}^2_{MESIM}$ and $\hat{\beta}_{MESIM}$ are the maximum likelihood estimators of the eigenvector and eigenvalue of $\mathbf{K}\mathbf{K}'$, respectively, and are asymptotically consistent and unbiased. The estimators of **K**, **L**, **H**, and Δ are $\hat{\mathbf{K}}, \hat{\mathbf{L}}, \hat{\mathbf{H}},$ and $\hat{\boldsymbol{\Delta}}$, respectively, so $\hat{\mathbf{K}} = \hat{\mathbf{L}}\hat{\boldsymbol{\Delta}}\hat{\mathbf{H}}'$.

INCORPORATING MULTITRAIT MULTIENVIRONMENT AND GENOMEWIDE MOLECULAR MARKER INFORMATION INTO SELECTION INDICES

Suppose one wishes to have a selection index that will facilitate selecting genotypes in environments while, at the same time, incorporating additional random variables represented by several MM. Such is the multitrait multienvironment genomewide molecular marker selection index (Y_M), which, according to Eq. 3 can be written as

$$Y_M = \boldsymbol{\beta}'_E\mathbf{p}_E + \boldsymbol{\beta}'_m\mathbf{m} = [\boldsymbol{\beta}'_E \quad \boldsymbol{\beta}'_m]\begin{bmatrix}\mathbf{p}_E \\ \mathbf{m}\end{bmatrix} = \boldsymbol{\beta}'_M\mathbf{p}_M \tag{7}$$

where $\boldsymbol{\beta}'_M = [\boldsymbol{\beta}'_E \quad \boldsymbol{\beta}'_m]$ and $\mathbf{p}_M = [\mathbf{p}'_E\,\mathbf{m}']$; β_E is a vector of weights for the phenotypic traits of the corresponding genotypic traits evaluated in various environments; β_m is the vector of molecular marker weights; $\mathbf{p}_E$ is the vector of the phenotypic values of the genotypes evaluated in various environments; and **m** is the molecular marker vector, where the homozygous genotyped molecular marker takes values of 1 and –1, and the heterozygous genotyped molecular marker takes a value of 0 for cases of segregating populations such as F_2 populations, or values of 1 and –1 for doubled haploids denoting the presence or absence of the molecular marker; a similar designation is given to the genotyped QTLs (see Appendix). Then the response to this selection index (Y_M) can be written as

$$R_M = k\sigma_{Z_M}\rho_{Y_M Z_M} = k\sigma_{Z_M}\frac{\boldsymbol{\theta}'_M\boldsymbol{\Sigma}\boldsymbol{\beta}_M}{\sqrt{\boldsymbol{\theta}'_M\boldsymbol{\Sigma}\boldsymbol{\theta}_M}\sqrt{\boldsymbol{\beta}'_M\boldsymbol{\Gamma}\boldsymbol{\beta}_M}} \tag{8}$$

where $\Gamma = \begin{bmatrix} \mathbf{P}_E & \mathbf{G}_M \\ \mathbf{G}'_M & \mathbf{M} \end{bmatrix}$, $\Sigma = \begin{bmatrix} \mathbf{G}_E & \mathbf{G}_M \\ \mathbf{G}'_M & \mathbf{M} \end{bmatrix}$; k has been defined as in Eq. 2; $\sigma^2_{Z_M} = \boldsymbol{\theta}'_M \boldsymbol{\Sigma} \boldsymbol{\theta}_M$ is the variance of the genomewide breeding value ($Z_M = \boldsymbol{\theta}'_E \mathbf{g}_E + \boldsymbol{\theta}'_m \mathbf{m}$); $\boldsymbol{\theta}'_M = [\boldsymbol{\theta}'_E \quad \boldsymbol{\theta}'_m]$ is the genomewide vector of economic weights (in the LW selection index, θ_m is a vector of zeros); $\boldsymbol{\beta}'_M \boldsymbol{\Gamma} \boldsymbol{\beta}_M$ is the variance of the selection index Y_M; and $\boldsymbol{\beta}'_M = [\boldsymbol{\beta}'_E \quad \boldsymbol{\beta}'_m]$ is the genomewide vector containing phenotypic values in the various environments (β_E) and molecular (β_m) weights. Γ and $\boldsymbol{\Sigma}$ are variance-covariance matrices comprising the phenotypic variance-covariance matrix of the genotypes evaluated in various environments ($\mathbf{P}_E$); the genotypic variance-covariance matrix of the genotypes evaluated in various environments ($\mathbf{G}_E$); the variance-covariance matrix of the molecular markers ($\mathbf{M}$), and the covariance matrix of the genotypic values with the molecular markers ($\mathbf{G}_M$).

The structure of matrices $\mathbf{P}_E$, $\mathbf{G}_E$, $\mathbf{M}$, and $\mathbf{G}_M$ are as follows

$$\mathbf{P}_E = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \cdots & \mathbf{P}_{1N_e} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \cdots & \mathbf{P}_{2N_e} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{P}_{N_e1} & \mathbf{P}_{N_e2} & \cdots & \mathbf{P}_{N_eN_e} \end{bmatrix}, \text{ and}$$

$$\mathbf{G}_E = \begin{bmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} & \cdots & \mathbf{G}_{1N_e} \\ \mathbf{G}_{21} & \mathbf{G}_{22} & \cdots & \mathbf{G}_{2N_e} \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{G}_{N_e1} & \mathbf{G}_{N_e2} & \cdots & \mathbf{G}_{N_eN_e} \end{bmatrix},$$

where $\mathbf{P}_{ee}$ and $\mathbf{G}_{ee}$ are the e^{th} phenotypic and genotypic variance-covariance matrices, respectively ($e = 1, 2, \ldots, N_e$, N_e = number of environments). The Appendix shows that, in the case of a doubled haploid population, matrix $\mathbf{M}$ can be written as

$$\mathbf{M} = \begin{bmatrix} 1 & (1-2r_{11}) & \cdots & (1-2r_{1N_m}) \\ (1-2r_{21}) & 1 & \cdots & (1-2r_{2N_m}) \\ \cdots & \cdots & \cdots & \cdots \\ (1-2r_{N_m1}) & (1-2r_{N_m2}) & \cdots & 1 \end{bmatrix},$$

where $(1-2r_{1m})$ denotes the covariance between any two MM, for example between MM 1 and MM m^{th} for $m = 1, 2, \ldots, N_m$ (where N_m = number of MM).

The structure of $\mathbf{G}_M$ depends on the way the genotypic values of the traits have been defined. Let g_{eit} be the genotypic value of the t^{th} trait of the i^{th} genotype evaluated in the e^{th} environment; assuming that the genotypes evaluated in different environments correspond to different random variables (*i.e.*, including combinations of traits and environments as phenotypic variables; in other words, phenotypic variable = trait-environment combination) (Falconer and Mackay 1997), then according to Lange and Whittaker (2001), g_{eit} can be written as

$$g_{eit} = \sum_{q=1}^{N_{QTLet}} x_{eiq} \alpha_{etq} \qquad (9)$$

$e = 1, 2, \ldots, N_e$, where x_{eiq} denotes the code value of the q^{th} QTL in the i^{th} genotype ($i = 1, 2, \ldots, n$, n = number of genotypes or sample size) in the e^{th} environment (in a double haploid population the values of x_{eiq} are 1 or –1); α_{etq} is the additive effect of the q^{th} QTL on the t^{th} trait ($t = 1, 2, \ldots, N_t$, N_t = number of traits under selection) in the e^{th} environment; and N_{QTLet} is the number of QTLs that affect the t^{th} trait in the e^{th} environment. The inconsistency of the values of α_{etq} in different environments indicates QTL × environment interaction, which leads to the detection of QTL effects in some environments but not in others (Beavis and Keim 1996; Crossa *et al.* 1999; Bernardo 2002).

Phenotypic values, according to Eq. 9, are modeled as

$$p_{eit} = g_{eit} + E_e + GE_{ie} + \varepsilon_{eit} \qquad (10)$$

where the first term of Eq. 10 is the same as in Eq. 9, E_e is the effects of the environments, GE_{ie} is the genotype × environment interaction effect, and ε_{eij} is the experimental error. For a double haploid population, assuming that the genotypes evaluated in different environments correspond to different random variables, the Appendix shows that $\mathbf{G}_M$ can be written as

$$\mathbf{G}_M = \begin{bmatrix} (1-2r_{11})\alpha_{11} & (1-2r_{12})\alpha_{12} & \cdots & (1-2r_{1N_m})\alpha_{1N_{QTL1}} \\ (1-2r_{21})\alpha_{21} & (1-2r_{22})\alpha_{22} & \cdots & (1-2r_{2N_m})\alpha_{2N_{QTL2}} \\ \cdots & \cdots & \cdots & \cdots \\ (1-2r_{N_t1})\alpha_{N_t1} & (1-2r_{N_t2})\alpha_{N_t2} & \cdots & (1-2r_{N_tN_m})\alpha_{N_tN_{QTLt}} \end{bmatrix},$$

where $(1-2r_{tm})\alpha_{tq}$ is the covariance between the code values of the m^{th} MM and the genotypic value of the t^{th} trait that is influenced by the genotypic additive effect of the q^{th} QTL ($t = 1, 2, \ldots, N_t$; $m = 1, 2, \ldots, N_m$;

$q = 1, 2, \ldots, N_{QTLt}$, where N_{QTL1} = number of QTLs that affect the genotype of trait 1, N_{QTL2} = number of QTLs that affect the genotype of trait 2, etc.). These results make it possible to develop LW, KN$_{\text{GW}}$, MESIM$_{\text{GW}}$, and RESIM$_{\text{GW}}$.

Multitrait multienvironment genomewide molecular marker indices of Lange-Whittaker (LW): In the multitrait multienvironment genomewide molecular marker selection index of Lange and Whittaker (2001), the vector $\beta_{LW} = \Gamma^{-1}\Sigma\theta_M$ allows constructing the genomewide molecular selection index $Y_{LW} = \boldsymbol{\beta}'_{LW}\mathbf{p}_M$ which has maximum correlation with $Z_M = \boldsymbol{\theta}'_E\mathbf{g}_E + \boldsymbol{\theta}'_m\mathbf{m}$; the maximized genomewide selection response can be written as $R_{LW} = k\sqrt{\boldsymbol{\beta}'_{LW}\boldsymbol{\Gamma}\boldsymbol{\beta}_{LW}}$. The estimators of β_{LW} ($\hat{\boldsymbol{\beta}}_{LW}$) for various traits in various environments are obtained directly from the estimators of $\mathbf{P}_E$, $\mathbf{G}_E$, $\mathbf{M}$, $\mathbf{G}_M$, ($\hat{\mathbf{P}}_E$, $\hat{\mathbf{G}}_E$, $\hat{\mathbf{M}}$, and $\hat{\mathbf{G}}_M$), and from the vector θ_M.

The restrictive multitrait multienvironment genomewide molecular marker selection index of Kempthorne-Nordskog (KN$_{\text{GW}}$): The development of the KN$_{\text{GW}}$ is direct. In this case, KN$_{\text{GW}}$ RSI is defined as $\boldsymbol{\beta}_{KN_{GW}} = \mathbf{A}_{GW}\beta_{LW}$, where $\mathbf{A}_{GW}$, = [$\mathbf{I} - \boldsymbol{\Gamma}^{-1}\mathbf{C}_{GW}(\mathbf{C}'_{GW}\boldsymbol{\Gamma}^{-1}\mathbf{C}_{GW})^{-1}\mathbf{C}'_{GW}$], Γ^{-1} is the inverse matrix of $\mathbf{C}_{GW} = \boldsymbol{\Sigma}\mathbf{F}$, and $\beta_{LW} = \Gamma^{-1}\Sigma\theta_M$. Thus, in the context of the KN$_{\text{GW}}$ RSI, the $Y_{KN_{GW}}$ that maximizes $\rho_{Y_M Z_M}$ is $Y_{KN_{GW}} = \boldsymbol{\beta}'_{KN_{GW}}\mathbf{p}$, and the maximized selection response is $R_{KN_{GW}} = k\sqrt{\boldsymbol{\beta}'_{KN_{GW}}\boldsymbol{\Gamma}\boldsymbol{\beta}_{KN_{GW}}}$.

The multitrait multienvironment genomewide molecular marker eigen selection index method (MESIM$_{\text{GW}}$): In MESIM$_{\text{GW}}$, β_M and Y_M can be obtained using the estimation procedure described by Cerón-Rojas *et al.* (2008b) by means of the singular value decomposition theory. In MESIM$_{\text{GW}}$ it is necessary to maximize

$$\Pi = (\boldsymbol{\theta}'_M\boldsymbol{\Sigma}\boldsymbol{\beta}_M)^2 - \lambda(\boldsymbol{\beta}'_M\boldsymbol{\Gamma}\boldsymbol{\beta}_M - 1) - \omega(\boldsymbol{\theta}'_M\boldsymbol{\Sigma}\boldsymbol{\theta}_M - 1)$$

with respect to β_M, θ_M, λ, and ω, where β_M is the vector of MESIM$_{\text{GW}}$ coefficients, θ_M is the vector of economic weights, λ and ω are Lagrange multipliers, and $\boldsymbol{\beta}'_M\boldsymbol{\Gamma}\boldsymbol{\beta}_M = 1$ and $\boldsymbol{\theta}'_M\boldsymbol{\Sigma}\boldsymbol{\theta}_M = 1$ are restrictions imposed when maximizing $\rho^2_{Y_M Z_M}$. In MESIM$_{\text{GW}}$ it is assumed that θ_M is not a vector of constants.

When $\boldsymbol{\Pi}$ is derived with respect to $\boldsymbol{\beta}_M$ and θ_M, the result is

$$(\mathbf{Q} - \lambda\mathbf{I})\boldsymbol{\beta}_{\mathrm{M}} = \mathbf{0} \tag{11}$$

where $\mathbf{Q} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Sigma}$. Thus, for MESIM$_{\text{GW}}$, the value that maximizes $\rho^2_{Y_M Z_M}$ is the first eigenvalue (λ) of matrix $\mathbf{Q}$, and the vector that allows constructing Y_M (with maximum correlation with $Z_M = \boldsymbol{\theta}'_E\mathbf{g}_E + \boldsymbol{\theta}'_m\mathbf{m}$) is the first eigenvector of matrix $\mathbf{Q}(\boldsymbol{\beta}_M)$; the maximized selection response can be written as $R_M = k\sqrt{\lambda}$. When information on the QTLs linked to the molecular markers is not incorporated into the selection index, but selection is conducted on various traits and environments simultaneously, then Eq. 11 can be written as $(\mathbf{P}_E^{-1}\mathbf{G}_E - \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{0}$ and $\mathbf{Q} = \mathbf{P}_E^{-1}\mathbf{G}_E$; when selection is conducted in one environment, $(\mathbf{P}^{-1}\mathbf{G} - \lambda\mathbf{I})$ $\boldsymbol{\beta} = \mathbf{0}$ and then $\mathbf{Q} = \mathbf{P}^{-1}\mathbf{G}$ (where $\mathbf{G}$ and $\mathbf{P}$ are defined as in Eq. 2), from where the relationship among ESIM, MESIM, and MESIM$_{\text{GW}}$ is clear.

Using singular value decomposition, $\mathbf{Q}$ can be written as

$$\mathbf{Q} = \mathbf{UDV}'$$

where the columns of matrix $\mathbf{U}(\mathbf{U}'\mathbf{U} = \mathbf{I})$ are the left singular vector of $\mathbf{Q}$, and the columns of matrix $\mathbf{V}(\mathbf{V}'\mathbf{V} = \mathbf{I})$ are the right singular vector of $\mathbf{Q}$; $\mathbf{D}$ is a diagonal matrix with singular values of $\mathbf{QQ}'$ or $\mathbf{Q}'\mathbf{Q}$. The estimators of $\lambda = \lambda_{MESIM_{GW}}$ and $\beta_M = \boldsymbol{\beta}_{MESIM_{GW}}$ are obtained from $\hat{\mathbf{Q}}\hat{\mathbf{Q}}'$, such that

$$(\hat{\mathbf{Q}}\hat{\mathbf{Q}}' - \hat{\lambda}^2_{MESIM_{GW}}\mathbf{I})\hat{\boldsymbol{\beta}}_{MESIM_{GW}} = \mathbf{0};$$

$\hat{\lambda}^2_{MESIM_{GW}}$ and $\hat{\boldsymbol{\beta}}_{MESIM_{GW}}$ are the maximum likelihood estimators of the eigenvector and eigenvalue of $\mathbf{QQ}'$, respectively, and are asymptotically consistent and unbiased. The estimators of $\mathbf{Q}$, $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{D}$ are $\hat{\mathbf{Q}}$, $\hat{\mathbf{U}}$, $\hat{\mathbf{V}}$, and $\hat{\mathbf{D}}$, respectively, so $\hat{\mathbf{Q}} = \hat{\mathbf{U}}\hat{\mathbf{D}}\hat{\mathbf{V}}'$. These results allow estimating Y_M as $\hat{Y}_M = \hat{\boldsymbol{\beta}}'_{MESIM_{GW}}\mathbf{p}_M$.

The restrictive multitrait multienvironment genomewide molecular marker eigen selection index method (RESIM$_{GW}$) : As in the case of KN$_{GW}$, developing RESIM$_{GW}$ is straightforward. The first eigenvalue ($\lambda_{RESIM_{GW}}$) that maximizes $\rho^2_{Y_M Z_M}$ under the three restrictions, $\boldsymbol{\beta}'_M \boldsymbol{\Gamma} \boldsymbol{\beta}_M = 1$, $\boldsymbol{\theta}'_M \boldsymbol{\Sigma} \boldsymbol{\theta}_M = 1$, and $\boldsymbol{\beta}'_M \mathbf{C}_{GW} = \mathbf{0}$, and the vector ($\boldsymbol{\beta}_{RESIM_{GW}}$) that allows constructing $Y_M = \boldsymbol{\beta}'_M \mathbf{p}$, which has maximum correlation with Z_M = $\boldsymbol{\theta}'_E \mathbf{g}_E + \boldsymbol{\theta}'_m \mathbf{m}$, provide the solution to the equation

$$(\mathbf{Q}_{RGW} - \lambda_{RESIM_{GW}} \mathbf{I}) \boldsymbol{\beta}_{RESIM_{GW}} = \mathbf{0}$$

where

$$\mathbf{Q}_{RGW} = [\mathbf{I} - \boldsymbol{\Gamma}^{-1} \mathbf{C}_{GW} (\mathbf{C}'_{GW} \boldsymbol{\Gamma}^{-1} \mathbf{C}_{GW})^{-1} \mathbf{C}'_{GW}] \boldsymbol{\Gamma}^{-1} \boldsymbol{\Sigma},$$

and $\mathbf{C}_{GW} = \boldsymbol{\Sigma}\mathbf{F}$.

SAMPLING PROPERTIES OF THE LANGE-WHITTAKER MULTITRAIT MULTIENVIRONMENT GENOMEWIDE MOLECULAR MARKER ESTIMATOR ($\hat{\boldsymbol{\beta}}_{LW}$) IN THE CONTEXT OF THE MESIM$_{GW}$

Sampling properties of the estimator $\hat{\boldsymbol{\beta}}_{LW}$ are important because they make it possible to determine how near or how far $\hat{\boldsymbol{\beta}}_{LW}$ is from the population parameter, $\boldsymbol{\beta}_{LW}$. In addition, this facilitates obtaining information on the selection index, Y_{LW}. Assuming that the variance-covariance matrix estimators of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ ($\hat{\boldsymbol{\Gamma}}$ and $\hat{\boldsymbol{\Sigma}}$) are independent, then the expectation of $\hat{\boldsymbol{\beta}}_{LW}$ = $\hat{\boldsymbol{\Gamma}}^{-1} \hat{\boldsymbol{\Sigma}} \boldsymbol{\theta}_M$ [$E(\hat{\boldsymbol{\beta}}_{LW})$] is relatively easy to obtain, since $\boldsymbol{\theta}_M$ is a vector of constants. However, the variance of $\hat{\boldsymbol{\beta}}_{LW}$ = [$Var(\hat{\boldsymbol{\beta}}_{LW})$], even in the unlikely case that $\hat{\boldsymbol{\Gamma}}$ and $\hat{\Sigma}$ are independent, is not simple to compute. It is evident that $\hat{\boldsymbol{\Gamma}}$ and $\hat{\Sigma}$ are not independent, and this may be the main reason why Harris (1964) and Hayes and Hill (1980) were not successful in determining the statistical sampling properties of the Smith (1936) selection index. Based on the sampling properties of $\hat{\boldsymbol{\beta}}_{MESIM_{GW}}$, we propose a method for finding the sampling properties of $\hat{\boldsymbol{\beta}}_{LW}$. The following procedure is also valid in the context of the restrictive multitrait multienvironment genomewide molecular marker selection index, KN$_{GW}$ and RESIM$_{GW}$, as well as the MESIM and LT selection indices applied to the case of multitrait multienvironment molecular markers, ESIM, and the Smith selection index.

For simplicity, assume that $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, …, $\boldsymbol{\beta}_{N_t+N_m}$ are the eigenvectors of $\mathbf{QQ}'$ from MESIM$_{GW}$; then, due to the Bessel inequality (Rao 2002),

$$\left|\boldsymbol{\beta}'_1 \boldsymbol{\beta}_{LW}\right|^2 + \left|\boldsymbol{\beta}'_2 \boldsymbol{\beta}_{LW}\right|^2 + \ldots + \left|\boldsymbol{\beta}'_{N_t+N_m} \boldsymbol{\beta}_{LW}\right|^2 \le \left\|\boldsymbol{\beta}_{LW}\right\|^2 \quad (12)$$

in which $\left\|\boldsymbol{\beta}_{LW}\right\| = \sqrt{\boldsymbol{\beta}'_{LW} \boldsymbol{\beta}_{LW}}$ denotes the Euclidean norm of $\boldsymbol{\beta}_{LW}$ and $\left|\boldsymbol{\beta}'_d \boldsymbol{\beta}_{LW}\right|^2$ (d = 1, 2, …, N_t + N_m, N_t + N_m = number of traits under selection (N_t) plus number of MM (N_m)) denotes the square of the absolute value of the scalar product of the d^{th} eigenvector of $\mathbf{QQ}'$ and $\boldsymbol{\beta}_{LW}$. Equality in Eq. (12) will occur when $\boldsymbol{\beta}_{LW}$ is in the subspace generated by the eigenvectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, …, $\boldsymbol{\beta}_{N_t+N_m}$ of $\mathbf{QQ}'$ (Rao 2002). Suppose that $\boldsymbol{\beta}_{LW}$ is in the subspace generated by the eigenvectors $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, …, $\boldsymbol{\beta}_{N_t+N_m}$; then, since the eigenvectors of $\mathbf{QQ}'$ form a base (Rao 2002), they are linearly independent and $\boldsymbol{\beta}_{LW}$ can be written as a linear combination of the eigenvectors of $\mathbf{QQ}'$, that is

$$\boldsymbol{\beta}_{LW} = a_1 \boldsymbol{\beta}_1 + a_2 \boldsymbol{\beta}_2 + \ldots + a_{N_t+N_m} \boldsymbol{\beta}_{N_t+N_m} = \mathbf{Ua} \quad (13)$$

where the coefficients a_d (d = 1, 2, …, N_t + N_m) denote scalars, $\mathbf{U}$($\mathbf{U}'\mathbf{U} = \mathbf{I}$) is a matrix with the eigenvectors of $\mathbf{QQ}'$, and $\mathbf{a}$ is a vector of a_d's. Eq. (13) is given in Cerón-Rojas *et al.* (2008a).

Note that $\mathbf{U}'\boldsymbol{\beta}_{LW} = \mathbf{a}$, *i.e.*, $a_d = \boldsymbol{\beta}'_d \boldsymbol{\beta}_{LW} = \boldsymbol{\beta}'_{LW} \boldsymbol{\beta}_d$, from where Eq. (13) can be written as

$$\begin{aligned}\boldsymbol{\beta}_{LW} = {} & (\boldsymbol{\beta}'_{LW} \boldsymbol{\beta}_1) \boldsymbol{\beta}_1 + (\boldsymbol{\beta}'_{LW} \boldsymbol{\beta}_2) \boldsymbol{\beta}_2 + \ldots \\ & + (\boldsymbol{\beta}'_{LW} \boldsymbol{\beta}_{N_t+N_m}) \boldsymbol{\beta}_{N_t+N_m} \qquad (14)\end{aligned}$$

The importance of Eq. (13) is that it represents a unique linear combination. To see this, suppose that another linear combination, such as $\mathbf{Ua} = \mathbf{Ub}$, exists; then $\mathbf{U}(\mathbf{a} - \mathbf{b}) = \mathbf{0}$; because the eigenvectors of $\mathbf{U}$ are linearly independent and different from $\mathbf{0}$, $(\mathbf{a} - \mathbf{b}) = \mathbf{0}$, from where $\mathbf{a} = \mathbf{b}$. Hence, the linear combination of Eq. (13) is unique. It is possible to show that the estimator of the vector $\mathbf{a}$ ($\hat{\mathbf{a}}$) is a least square estimator. Let

$$\begin{aligned}\delta &= (\boldsymbol{\beta}_{LW} - \mathbf{Ua})' (\boldsymbol{\beta}_{LW} - \mathbf{Ua}) \\ &= \boldsymbol{\beta}'_{LW} \boldsymbol{\beta}_{LW} + \mathbf{a}'\mathbf{U}'\mathbf{Ua} - 2\mathbf{a}'\mathbf{U}'\boldsymbol{\beta}_{LW}\end{aligned}$$

Because $\mathbf{U'U} = \mathbf{I}$, when deriving δ with respect to $\mathbf{a}$, we get

$$\frac{\partial}{\partial \mathbf{a}}\delta = 2\mathbf{Ia} - 2\mathbf{U}'\boldsymbol{\beta}_{LW}$$

When this is equal to the null vector, $\hat{\mathbf{a}} = \hat{\mathbf{U}}'\hat{\boldsymbol{\beta}}_{LW}$, and because $\frac{\partial^2}{\partial \mathbf{a}^2}\delta = 2\mathbf{I}$, then $\hat{\mathbf{a}} = \hat{\mathbf{U}}'\hat{\boldsymbol{\beta}}_M$ effectively minimizes the distance between $\boldsymbol{\beta}_{LW}$ and $\mathbf{Ua}$. Therefore, $\hat{\mathbf{a}}$ is an unbiased estimator with minimum variance (characteristics of the least square estimators).

The previous results indicate that the estimator of $\boldsymbol{\beta}_{LW}$ ($\hat{\boldsymbol{\beta}}_{LW}$) can be written as a linear combination of the eigenvector estimators of $\mathbf{QQ'}$, *i.e.*,

$$\hat{\boldsymbol{\beta}}_{LW} = \hat{a}_1\hat{\boldsymbol{\beta}}_1 + \hat{a}_2\hat{\boldsymbol{\beta}}_2 + \ldots + \hat{a}_{N_t+N_m}\hat{\boldsymbol{\beta}}_{N_t+N_m} = \hat{\mathbf{U}}\hat{\mathbf{a}} \quad (15)$$

where $\hat{a}_d$ is the d^{th} (d = 1, 2, …, $N_t + N_m$, $N_t + N_m$ = number of traits under selection (N_t) plus number of MM (N_m)) element of $\hat{\mathbf{a}}$.

Note that $\hat{\mathbf{a}}$ is an unbiased estimator with minimum variance and that each $\hat{a}_d$ is a scalar whereas each $\hat{\boldsymbol{\beta}}_d$ is a vector. Suppose that $\hat{a}_d$ and $\hat{\boldsymbol{\beta}}_d$ are independent, d = 1, 2, …, $N_t + N_m$, since the sampling properties of $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, …, $\boldsymbol{\beta}_{N_t+N_m}$ are known (Mardia *et al.* 1982; Anderson 2003), then the sampling properties of $\hat{\boldsymbol{\beta}}_{LW}$ are also known. In MESIM$_{GW}$, asymptotically $E(\hat{\boldsymbol{\beta}}_d) = \boldsymbol{\beta}_d$ and for $d \neq f$, $Var(\hat{\boldsymbol{\beta}}_d) \approx \frac{1}{n}\sum_{f=1}^{N_t+N_m}\frac{\lambda_d\lambda_f}{(\lambda_d-\lambda_f)^2}\boldsymbol{\beta}_f\boldsymbol{\beta}'_f$, and $Cov(\hat{\boldsymbol{\beta}}_d,\hat{\boldsymbol{\beta}}_f) \approx -\frac{\lambda_d\lambda_f}{n(\lambda_d-\lambda_f)^2}\boldsymbol{\beta}_d\boldsymbol{\beta}'_f$ where d, f = 1, 2, …, $N_t + N_m$, and n is the number of genotypes or sample size.

SIMULATED AND REAL DATA

Simulated data: We used the genetic and breeding simulation tool of QuLine (previously called QuCim) (Wang *et al.* 2003, 2004) to simulate genotypes from a population with the aim of assessing the theoretical and practical results from the MESIM$_{GW}$, LW, KN$_{GW}$, and RESIM$_{GW}$ selection indices applied to the case of multitrait multienvironment genomewide molecular markers, and MESIM and LT selection indices applied to the case of multitrait multienvironment molecular markers. To simulate the genotypic and phenotypic values of individuals in a breeding population, a genetic model (which is called gene and environment (GE) system in QuLine) needs to be defined first. The information required for defining a GE system includes number of genes (or QTLs), gene effect for each trait (including additive, dominance, and epistasis), linkage among the genes in one chromosome, trait heritability, etc.

On the other hand, a breeding strategy to generate various breeding populations needs to be defined as well. By defining breeding strategy, QuLine translates the complicated breeding process into one the computer can understand and simulate. QuLine allows several breeding strategies to be defined simultaneously; they are contained in one input file. The program then makes the same virtual crosses for all the defined strategies in the first breeding cycle. A breeding strategy in QuLine is defined as all the crossing, seed propagation, and selection activities in an entire breeding cycle. A breeding cycle begins with crossing and ends at the generation when the selected advanced lines are returned to the crossing block as new parents. Selection methods that can be simulated in QuLine include mass selection, pedigree system, bulk population system, backcross breeding, top-cross breeding, doubled haploid breeding, marker-assisted selection for one trait, and many combinations and modifications of these. The simulator provides the true genotypic value for each genotype in the population, as well as the phenotypic value of the traits under study.

Using a sample of 240 genotypes and 125 molecular markers from a real F$_3$ maize population data set, selection was made based on five traits in three environments (15 variables); the 15 variables were considered simultaneously.

Generating the simulated doubled haploid population for selection: We followed the procedures described by Zhang and Smith (1992), in which the additive effects of the QTLs required for defining a GE system in QuLine were obtained from a Normal distribution of gene effects (both positive and negative) that contribute to total additive genetic variance. The simulated traits were female flowering time (FFL) (days) and grain yield (GY) (grams per plot) measured in three different environments for a maize population of 10 chromosomes. A total of 460 MM were distributed every 5 cM over the 10 chromosomes. Also, 49 QTLs for FFL were randomly distributed, with a total of 13 QTLs for each environment and 120 QTLs for GY with 40 QTLs in each environment. The data were used to generate 500 doubled haploid genotypes that form the reference population (cycle 0). The two traits were considered simultaneously in the three environments, together with the 460 MM. Using 10% (k=1.755) selection pressure, 50 genotypes were selected under MESIM_{GW}, LW, KN_{GW}, and RESIM_{GW} multitrait multienvironment genomewide MM selection index, and marker-assisted selection indices MESIM and LT were applied to the case of multitrait multienvironment MM. The 50 selected doubled haploids were then crossed in diallel fashion, and a new population of 500 doubled haploids was generated. This was repeated during five selection cycles for the two traits in the three environments using all 460 MM.

The efficiency of the indices was compared using the true mean genotypic value and the regression of the mean genotypic value of the selected genotypes on the selection cycles. We used phenotypic, genotypic, and molecular marker variance-covariance matrices for estimating the singular vectors and the singular values for MESIM_{GW}, RESIM_{GW}, and MESIM, as well as the weights of the coefficients of LW, KN_{GW}, and LT. The true genotypic values of each individual are given directly and do not need to be estimated from the data. The RESIM_{GW} and KN_{GW} were applied for fixing FFL and selecting just for GY using all the available markers.

Sign of the coefficients and economic weights of the selection indices: When using MESIM_{GW}, and MESIM, it is often necessary to change the sign of the coefficients of the first singular eigenvector in order to select the genotypes according to the desired genetic advance, that is, for trait FFL the sign is always negative (decreasing the mean genotypic value), whereas for GY, the sign is always positive (increasing the mean genotypic value). However, when MESIM_{GW} and RESIM_{GW} are used, the number of QTLs affecting the trait and the molecular markers linked to the QTLs are unknown, so the sign (direction) of the molecular markers on the MESIM_{GW} cannot be modified as in the case of MESIM, where it is possible to modify the direction of the coefficients of molecular scores.

As for the economic weights of the LW, LT, and KN_{GW} selection indices, they were assigned following Smith *et al.* (1981). One set of economic weights had coefficients of 1 or –1, and the other had the heritability of each trait multiplied by 1 or –1, depending on the trait. Therefore, for FFL and GY, the first set of economic weights was –1 and 1, respectively, in the three environments. The other set of economic weights is the heritability of each trait in each environment multiplied by –1 (for FFL) and by 1 (for GY); thus the heritabilities of the two traits in the first environment were h^2_{FFL1} =0.385 and h^2_{GY1} = 0.260; in the second environment, the heritabilities were h^2_{FFL2} = 0.579 and h^2_{GY2} = 0.506, and in the third environment the heritabilities were h^2_{FFL3} = 0.653 and h^2_{GY3} = 0.200; all economic weights of the molecular markers were equal to zero. The LW, KN_{GW}, and LT selection indices are denoted as LW1, KN1_{GW}, and LT1 when the economic weights are –1 and 1, and as LW2, KN2_{GW}, and LT2 when heritabilities are used as economic weights.

The two traits in the three environments, as well as all the molecular markers, were simultaneously considered for the selection indices MESIM_{GW}, LW, KN_{GW}, RESIM_{GW}, MESIM, and LT.

Real data: A real maize population with 240 F_3 genotypes and 125 molecular markers were used. This data set gave rise to the simulated data set described above. Selection was based on five traits evaluated in three environments. The 240 families were planted in the field using an incomplete block design with two replications. The signs of the scores in MESIM_{GW}, RESIM_{GW}, MESIM, and of the economic weights for LW, KN_{GW}, and LT were similar to those used for the simulated data, that is, in the LW selection index, one set had economic weights –1, –1, –1, –1, and 1 for male flowering time (MFL) (days), female flowering time (FFL) (days), ear height (EHT) (grams per plot), respectively, in the three environments, whereas the second set of economic weights comprised the heritability of the trait in each environment, as h^2_{MFL1}= 0.39, h^2_{FFL1}= 0.38, h^2_{FHT1} = 0.31, h^2_{PHT1}=0.19, h^2_{GY1} =

0.26, h^2_{MFL2} = 0.44, h^2_{FFL2} = 0.58, h^2_{EHT2} = 0.43, h^2_{PHT2} = 0.29, h^2_{GY2} = 0.51, h^2_{MFL3} = 0.75, h^2_{FFL3} = 0.65, h^2_{EHT3} = 0.48, h^2_{PHT3} = 0.39, and h^2_{GY3} = 0.18. All economic weights of the molecular markers were equal to zero. All five traits were simultaneously selected in three environments under MESIM$_{GW}$, LW, KN$_{GW}$, RESIM$_{GW}$, MESIM, and LT selection indices. In this case, the RESIM$_{GW}$ and KN$_{GW}$ were applied, the traits MFL and FFL were fixed in three different environments, and traits EHT, PHT, and GY were selected using all the available markers.

RESULTS AND DISCUSSION

Simulated data: Shown in Table 1 are the genotypic means of the lines selected using MESIM$_{GW}$, LW1, LW2, KN1$_{GW}$, KN2$_{GW}$, RESIM$_{GW}$, MESIM, LT1, and LT2 selection indices, when selection was practiced on two traits and three environments simultaneously during five selection cycles. In general, the average gains per selection cycle of the genomewide selection indices MESIM$_{GW}$, KN1$_{GW}$, KN2$_{GW}$, RESIM$_{GW}$, LW1, and LW2 were higher than the average gains per selection cycle of MESIM, LT1, and LT2 for trait GY in the three environments. These results agree with those obtained by Lange and Whittaker (2001), where genomewide was superior to marker-assisted selection. However, for trait FFL the average gains per selection cycle of MESIM and LT were better than the average gains per selection cycle of MESIM$_{GW}$, KN$_{GW}$, RESIM$_{GW}$, and LW in environments 2 and 3. Apparently, selection based on genomewide MM of traits with low heritability such as GY (h^2_{GY1} = 0.260; h^2_{GY2} = 0.506, and h^2_{GY3} = 0.200) is more effective than in traits with higher heritability such as FFL (h^2_{FFL1} = 0.385, h^2_{FFL2} = 0.579, and h^2_{FFL3} = 0.653).

Results of specific comparisons among various selection indices indicate that MESIM$_{GW}$ was superior to MESIM four out of six times, LW1 gave higher genetic gains than LT1 four out of six times, and five out of six times LW2 was more effective than LT2 in selecting the best lines. Within genomewide selections indices, results based on the average genetic gain per selection cycle indicated that MESIM $_{GW}$ had a slight advantage over LW1 and LW2 for GY. Average gains per selection cycle in environment 1 for GY were higher for MESIM$_{GW}$ (51.8 grams per plot) than those obtained using LW1 (49.6 grams per plot) and LW2 (46.7 grams per plot) (Table 1), but RESIM$_{GW}$ had the highest average genetic gains per selection cycle, 53.2 grams per plot for GY in environment 1. However, for FFL (–6.0) in environment 1, the opposite was true for LW1 (–6.6 days) but not for LW2 (–4.5 days). In environment 2, average gains in GY per selection cycle for MESIM$_{GW}$ were superior to those obtained using LW1 and LW2 (MESIM$_{GW}$ = 46.7 grams per plot versus LW1 = 42.2 grams per plot and LW2 = 39.9 grams per plot). In environment 3, average gains in GY from LW1 were the highest (23.4 grams per plot), as compared with MESIM$_{GW}$ (20.7 grams per plot) and LW2 (17.4 grams per plot) (Table 1). For FFL, LW1 and LW2 are more effective than MESIM$_{GW}$ for selecting early lines (*i.e.*, low FFL). When FFL was fixed and GY was selected using all the available markers, KN1$_{GW}$ and KN2 $_{GW}$ were more effective in environments 2 and 3 than RESIM$_{GW}$; however, RESIM$_{GW}$ was the best in environment 1.

Figs. 1 and 2 show the genotypic means for GY3 (GY in environment 3) and FFL2 (FFL in environment 2) for five selection cycles when genotypes were selected using MESIM$_{GW}$, LW1, LW2, KN1$_{GW}$, KN2$_{GW}$, RESIM$_{GW}$, MESIM, LT1, and LT2 selection indices. The effectiveness of the three genomewide selection indices (MESIM$_{GW}$, LW, and KN$_{GW}$) for increasing GY is clearly shown in Fig. 1. At the end of cycle 5, LW1 was the index that accumulated the highest selection gains, followed by KN1$_{GW}$, MESIM$_{GW}$, and RESIM$_{GW}$ (similar to MESIM). For FFL2 (Fig. 2) at the end of cycle 5, MESIM was the best selection index in terms of decreasing the maturity of the lines, followed by the marker-assisted selection indices LT1 and LT2, and LW2 and MESING$_{GW}$. The genomewide selection index LW2 was the best in cycle 4, but it did not show good gains after the last selection cycle. Note that, as expected, selection gains of the restrictive selection indices KN1$_{GW}$, KN2$_{GW}$, and RMESIM$_{GW}$ fluctuated around the original mean (cycle 0), since these selection indices do not change the trait FFL. Results from the simulation data did show the effectiveness of MESIM$_{GW}$, LW, KN$_{GW}$, and RESIM$_{GW}$ for improving the traits under selection in a multitrait multienvironment genomewide framework. For most of the trait-environment combinations, LW and MESIM$_{GW}$ were similar but better than MESIM and LT selection indices in terms of average genetic gains.

Table 1. Mean genotypic values of genotypes selected using MESIM_{GW}, Lange-Whittaker (LW1 and LW2), RESIM_{GW}, and Kempthorne-Nordskog (KN1_{GW} and KN2_{GW}), MESIM and Lande-Thompson (LT1 and LT2) for days to female flowering (FFL) and grain yield (GY) in three environments simultaneously for five cycles using imulated data. The signs and economic weights of the selection indices for each trait are shown in parentheses.

	MESIM_{GW}						LW1						LW1					
	Environment 1		Environment 2		Environment 3		Environment 1		Environment 2		Environment 3		Environment 1		Environment 2		Environment 3	
Selection cycle	FFL1 (–)	GY1 (+)	FFL2 (–)	GY2 (+)	FFL3 (–)	GY3 (+)	FFL1 (–1)	GY1 (+1)	FFL2 (–1)	GY2 (+1)	FFL3 (–1)	GY3 (+1)	FFL1 (–0.38)	GY1 (+0.26)	FFL2 (–0.58)	GY2 (+0.51)	FFL3 (–0.65)	GY3 (+0.20)
0	115.6	543.4	120.1	598.9	125.2	601.2	115.6	543.4	120.1	598.9	125.2	601.2	115.6	543.4	120.1	598.9	125.2	601.2
1	114.8	570.3	119.8	614.0	123.5	604.2	111.5	593.0	119.6	634.6	122.7	619.9	112.6	583.4	119.5	628.7	123.1	616.1
2	102.8	671.7	116.0	692.7	124.6	662.5	101.2	669.2	119.3	701.3	120.9	673.7	103.7	678.8	116.2	704.6	114.8	646.7
3	99.3	728.1	116.1	736.1	125.8	674.3	95.0	720.2	117.4	742.6	121.1	698.9	99.2	733.5	115.1	745.3	112.3	665.9
4	94.4	758.5	113.4	786.6	130.3	683.8	90.0	753.0	115.5	776.6	118.5	702.6	99.1	759.7	111.4	777.0	106.6	674.4
5	86.3	781.6	112.5	813.5	133.0	695.8	83.4	784.2	116.6	800.3	117.5	710.6	93.4	753.4	112.6	780.9	121.4	683.8
Average gain per selection cycle	**–6.0**	**51.8**	**–1.6**	**46.7**	**1.7**	**20.7**	**–6.6**	**49.6**	**–0.9**	**42.1**	**–1.5**	**23.4**	**–4.5**	**46.7**	**–1.8**	**39.9**	**–2.0**	**17.4**
	MESIM						LT1						LT2					
	Environment 1		Environment 2		Environment 3		Environment 1		Environment 2		Environment 3		Environment 1		Environment 2		Environment 3	
Selection cycle	FFL1 (–)	GY1 (+)	FFL2 (–)	GY2 (+)	FFL3 (–)	GY3 (+)	FFL1 (–1)	GY1 (+1)	FFL2 (–1)	GY2 (+1)	FFL3 (–1)	GY3 (+1)	FFL1 (–0.38)	GY1 (+0.26)	FFL2 (–0.58)	GY2 (+0.51)	FFL3 (–0.65)	GY3 (+0.20)
0	115.6	543.4	120.1	598.9	125.2	601.2	115.6	543.4	120.1	598.9	125.2	601.2	115.6	543.4	120.1	598.9	125.2	601.2
1	103.8	609.0	120.0	665.5	121.6	626.2	104.8	600.7	120.0	663.7	119.3	629.2	104.8	595.0	118.6	670.5	118.8	621.7
2	99.4	655.8	116.1	698.9	122.3	648.2	101.1	668.4	118.6	713.4	120.8	653.5	100.2	659.5	118.1	710.8	119.1	646.1
3	93.7	690.3	112.7	718.9	119.1	671.6	95.3	690.0	115.8	732.3	117.2	658.3	98.0	680.7	117.1	738.7	112.8	650.7
4	94.3	719.1	111.8	747.8	120.0	688.0	93.0	715.2	114.6	751.8	115.7	667.4	97.0	699.7	114.1	770.1	109.4	660.8
5	92.0	729.3	110.1	780.3	118.7	685.4	90.2	741.0	111.5	777.3	111.3	675.7	95.9	719.3	111.5	790.1	109.0	671.4
Average gain per selection cycle	**–4.3**	**37.0**	**–2.2**	**33.5**	**–1.2**	**18.0**	**–4.8**	**38.7**	**–1.8**	**33.6**	**–2.4**	**14.1**	**–3.5**	**34.7**	**–1.7**	**36.6**	**-3.3**	**13.5**
	RESIM_{GW}						KN1_{GW}						KN2_{GW}					
	Environment 1		Environment 2		Environment 3		Environment 1		Environment 2		Environment 3		Environment 1		Environment 2		Environment 3	
Selection cycle	FFL1 (–)	GY1 (+)	FFL2 (–)	GY2 (+)	FFL3 (–)	GY3 (+)	FFL1 (–1)	GY1 (+1)	FFL2 (–1)	GY2 (+1)	FFL3 (–1)	GY3 (+1)	FFL1 (–0.38)	GY1 (+0.26)	FFL2 (–0.58)	GY2 (+0.51)	FFL3 (–0.65)	GY3 (+0.20)
0	115.6	543.4	120.1	598.9	125.2	601.2	115.6	543.4	120.1	598.9	125.2	601.2	115.6	543.4	120.1	598.9	125.2	601.2
1	117.5	547.3	120.2	604.6	122.8	614.1	115.3	594.3	121.1	625.0	125.1	623.1	115.6	597.2	120.9	645.6	125.5	616.5
2	115.4	661.0	120.4	652.5	125.0	641.7	117.9	675.4	119.9	675.9	125.1	663.9	117.7	661.2	120.9	709.9	125.7	643.4
3	115.6	707.3	119.6	702.2	126.1	643.8	114.7	725.1	119.9	718.2	126.1	680.4	114.7	711.2	120.3	758.7	128.7	645.6
4	120.4	751.1	120.5	729.4	126.1	669.0	115.8	759.1	120.1	752.8	127.3	699.2	111.5	747.6	121.2	793.1	128.4	657.3
5	118.8	784.5	119.2	761.3	128.7	686.5	114.4	778.0	121.2	780.3	125.6	709.3	108.9	776.6	121.7	809.2	129.9	675.9
Average gain per selection cycle	**0.7**	**53.2**	**–0.1**	**35.3**	**0.8**	**17.0**	**–0.2**	**49.1**	**0.1**	**38.1**	**0.3**	**22.4**	**–1.4**	**47.6**	**0.2**	**44.1**	**1.0**	**14.2**

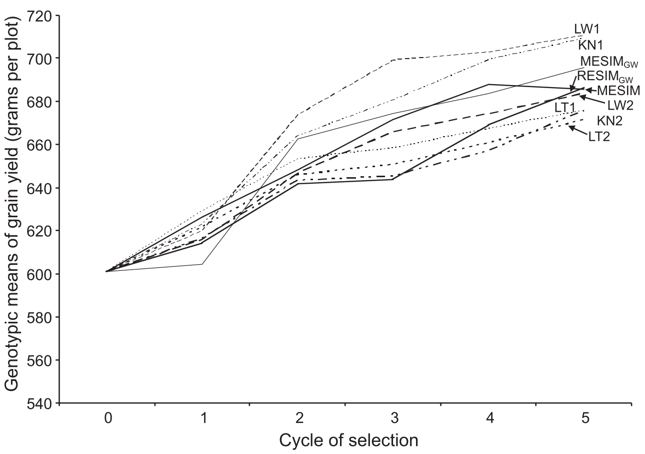

Fig. 1. Mean of the genotypic values of grain yield (GY) (grams per plot) in environment 3 of genotypes selected using MESIM$_{GW}$, Lange-Whittaker (LW1 and LW2), RESIM$_{GW}$, and Kempthorne-Nordskog (KN1$_{GW}$ and KN2$_{GW}$), multitrait multienvironment genomewide molecular marker selection index, and MESIM and Lande-Thompson (LT1 and LT2) multitrait multienvironment molecular marker selection index when two traits are selected in three environments simultaneously during five selection cycles using simulated data. The simultaneously selected traits were female flowering (FFL) and grain yield (GY). The economic weights used for female flowering (FFL) and grain yield (GY) under the Lange-Whittaker (LW1 and LW2) and Kempthorne-Nordskog (KN1$_{GW}$ and KN2$_{GW}$), multitrait multienvironment genomewide molecular marker selection index, and MESIM and Lande-Thompson (LT1 and LT2) multitrait multienvironment molecular marker selection index were –1 and 1, respectively, and the heritability of the corresponding traits.

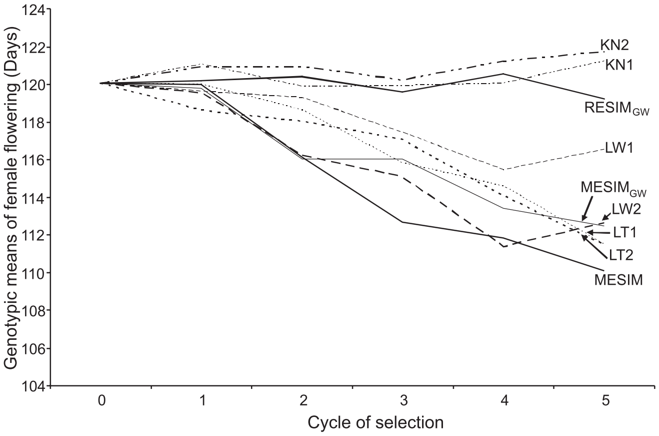

Fig. 2. Mean of the genotypic values of female flowering (FFL) (days) in environment 2 of genotypes selected using MESIM$_{GW}$, Lange-Whittaker (LW1 and LW2), RESIM$_{GW}$, and Kempthorne-Nordskog (KN1$_{GW}$ and KN2$_{GW}$), multitrait multienvironment genomewide molecular marker selection index, and MESIM and Lande-Thompson (LT1 and LT2) multitrait multienvironment molecular marker selection index when two traits are selected in three environments simultaneously during five selection cycles using simulated data. The restrictive trait was female flowering (FFL). The economic weights used for female flowering (FFL) and grain yield (GY) under the Lange-Whittaker (LW1 and LW2), and Kempthorne-Nordskog (KN1$_{GW}$ and KN2$_{GW}$), multitrait multienvironment genomewide molecular marker selection index, and MESIM and Lande-Thompson (LT1 and LT2) multitrait multienvironment molecular marker selection index were –1 and 1, respectively, and the heritability of the corresponding traits.

Real data: Table 2 shows the mean phenotypic values for one selection cycle of the 24 selected genotypes (10%) obtained using a real maize population with 240 F$_3$ genotypes and 125 molecular markers for MESIM$_{GW}$, LW (LW1 and LW2), MESIM, and LT (LT1 and LT2) when all five traits are selected in three different environments, and when RESIM$_{GW}$ and KN$_{GW}$ (KN1$_{GW}$ and KN2 $_{GW}$) were applied for fixing MFL and FFL in three different environments and for selecting for EHT, PHT, and GY using all the available markers. Note that MESIM$_{GW}$ was more efficient than LW for selecting shorter and earlier maize genotypes with higher grain production for all traits in all environments, except for PHT in environment 1 (PHT1) and for GY in environment 2 (GY2). These results indicate that when 15 traits were selected for simultaneously using all 125 available molecular markers, MESIM$_{GW}$ was better than LW in 13 of the 15 traits.

Results from the real data indicate that, at least for GY, MESIM$_{GW}$ would be more efficient than LW when trait heritability is low (Table 2). For example, heritabilities of GY in environments 1, 2, and 3 were $h^2_{GY1} = 0.26$, $h^2_{GY2} = 0.51$, and $h^2_{GY3} = 0.18$, respectively, and MESIM$_{GW}$ had a phenotypic mean of selected individuals higher than LW in environment 1 and environment 3, (Table 2). According to these results, the efficiency of the multitrait multienvironment genomewide molecular marker selection indices MESIM$_{GW}$ and LW depends on trait heritability.

Finally, it is worth noting that although MESIM$_{GW}$ may occasionally not be the selection index with the highest selection gains, it has the statistical properties of the principal components, which are easy to evaluate. In contrast, the statistical properties of the LW selection index are more difficult to assess. MESIM$_{GW}$ has some

Table 2. Mean phenotypic values of genotypes selected using MESIM_{GW}, Lange-Whittaker (LW1 and LW2), RESIM_{GW}, and Kempthorne-Nordskog (KN1_{GW} and KN2_{GW}), MESIM and Lande-Thompson (LT1 and LT2) for male flowering (MFL), female flowering (FFL), plant height (PHT), ear height (EHT), and grain yield (GY) in three environments from a real data set of a F_3 maize population using phenotypic, genotypic, and molecular marker variance-covariance matrices for one selection cycle. Heritability of each trait is shown in parentheses.

Genotypic means															
	Environment 1					Environment 2					Environment 3				
Selection Indices	MFL1 (0.39)	FFL1 (0.38)	EHT1 (0.31)	PHT1 (0.19)	GY1 (0.26)	MFL2 (0.44)	FFL2 (0.58)	EHT2 (0.43)	PHT2 (0.29)	GY2 (0.51)	MFL3 (0.75)	FFL3 (0.65)	EHT3 (0.48)	PHT3 (0.39)	GY3 (0.18)
MESIM_{GW}	98.8	98.5	75.0	141.8	166.7	98.1	98.1	72.1	131.6	109.3	102.1	103.4	71.0	116.1	144.0
LW1	100.8	100.7	78.1	141.5	164.9	98.8	98.9	76.5	134.4	100.8	102.9	104.7	73.0	113.8	81.8
LW2	100.5	100.4	78.5	141.7	131.3	98.5	98.3	76.1	133.6	111.4	103.1	105.0	71.5	111.4	70.5
MESIM	99.1	98.8	73.6	138.7	130.2	97.8	96.2	72.0	134.0	120.4	102.2	102.7	69.3	107.3	112.0
LT1	99.5	99.5	76.9	142.1	203.8	98.1	97.9	75.9	136.9	159.7	101.7	102.5	71.3	114.0	98.6
LT2	99.2	99.1	77.8	144.0	230.5	98.0	97.6	76.7	136.0	162.2	102.5	104.0	72.0	112.2	74.8
RESIM_{GW}	101.1	100.6	77.0	137.0	86.1	99.3	101.1	75.6	132.7	59.8	102.2	104.8	73.2	118.7	135.7
KN1_{GW}	101.5	101.5	79.7	140.7	124.9	100.7	98.8	80.1	135.4	73.5	103.4	104.8	72.7	112.5	70.6
KN2_{GW}	101.7	101.9	80.5	143.2	130.2	100.5	98.7	80.3	136.2	85.1	103.3	105.0	77.3	116.4	81.5
Original means	101.8	102.0	80.6	140.5	75.5	100.4	100.7	78.9	134.3	58.6	103.3	105.1	74.7	116.2	97.2

advantages over LW. First, it can be used to solve practical problems faced by breeders attempting to select plants or animals for the next generation when no estimates of economic weights are available. Even if economic weights are available, in practice it is very unlikely that they would maximize the derivative of $\boldsymbol{\theta}'_M \boldsymbol{\Sigma} \boldsymbol{\beta}_M$ with respect to $\boldsymbol{\beta}_M$ and to θ_M. Second, if two breeders are interested in improving, say, *n* traits, it is very unlikely that they would assign the same weights. Third, estimates of MESIM_{GW} have known statistical sampling properties that are easy to evaluate.

On the other hand, both MESIM_{GW} and LW have the main advantage of considering all possible types of cross products, *i.e.*, marker × marker, trait × trait in one environment, trait × trait in different environments, marker × trait in one environment, and marker × trait in different environments. Therefore, these selection indices implicitly consider the bi-genetic epistatic interaction networks that could potentially affect the expression of complex phenotypic traits of economic importance, such as grain yield in plants or meat and milk production in animals. Furthermore, MESIM_{GW} and LW consider possible epistatic interaction networks in complex *inter loci* interaction under different environmental conditions. Thus, it is expected that variability due to complicated interactions between epistatic networks × environments could be captured under the cross-products considered by MESIM_{GW} and LW.

In general, results indicate that MESIM was better than LT (LT1 and LT2), except for GY in environments 1 and 2, where LT2 was superior. Comparing RESIM_{GW} vs KN_{GW} (KN1_{GW} and KN2_{GW}), it is observed that for 11 variables RESIM_{GW} was better than KN_{GW} (Table 2). In general, when comparing MESIM_{GW}, MESIM, and RESIM_{GW} vs LW, LT, and KN_{GW}, results indicate that for 10 of the 15 traits the selection indices based on singular value decomposition were more efficient than LW, LT, and KN_{GW}.

Since singular value decomposition is a natural mathematical method for dimension reduction, MESIM_{GW} should be useful for direct use in marker-assisted recurrent selection. Although MESIM_{GW} is very straightforward, two difficulties are encountered when this method is used: (1) it is not possible to change the direction of the marker scores of the individuals of the first eigenvector, and (2) when applying MESIM_{GW} or LW, the user may have difficulty manipulating large variance-variance-covariance matrices with hundreds of thousands of cross products.

CONCLUSIONS

This research extended MESIM, RESIM (Cerón-Rojas *et al*. 2008 a, b), the Kempthorne and Nordskog (1959) restrictive selection index, and the Lange and Whittaker (2001) molecular selection indices to the case of a multitrait multienvironment genomewide molecular marker selection index. Results from real data showed that, in general, when several traits were selected in various environments simultaneously, $\mathrm{MESIM}_{\mathrm{GW}}$, MESIM, and $\mathrm{RESIM}_{\mathrm{GW}}$ increased the phenotypic means over the mean of individuals selected by the LW, LT, and $\mathrm{KN}_{\mathrm{GW}}$ selection indices. Results from simulated data did show some advantages of $\mathrm{MESIM}_{\mathrm{GW}}$, MESIM, and $\mathrm{MESIM}_{\mathrm{GW}}$ over LW, LT, and $\mathrm{KN}_{\mathrm{GW}}$ in terms of average genetic gains for some trait-environment combinations. However, LW, LT, and $\mathrm{KN}_{\mathrm{GW}}$ were sometimes superior to $\mathrm{MESIM}_{\mathrm{GW}}$, MESIM, and $\mathrm{RESIM}_{\mathrm{GW}}$.

One of the most important results of $\mathrm{MESIM}_{\mathrm{GW}}$, MESIM, and $\mathrm{RESIM}_{\mathrm{GW}}$ is that $\hat{\boldsymbol{\beta}}_{MESIM_{GW}}$, $\hat{\boldsymbol{\beta}}_{MESIM}$, and $\hat{\boldsymbol{\beta}}_{RESIM_{GW}}$ are the maximum likelihood estimators of $\boldsymbol{\beta}_{MESIM_{GW}}$, $\boldsymbol{\beta}_{MESIM}$, and $\boldsymbol{\beta}_{RESIM_{GW}}$, whereas $\hat{\boldsymbol{\beta}}_{LW}$, $\hat{\boldsymbol{\beta}}_{ps}$ (estimator of the molecular LT selection index), and $\hat{\boldsymbol{\beta}}_{KN_{GW}}$ are estimators of $\boldsymbol{\beta}_{LW}$, $\boldsymbol{\beta}_{ps}$, and $\boldsymbol{\beta}_{KN_{GW}}$, whose sampling properties are difficult to evaluate. $\mathrm{MESIM}_{\mathrm{GW}}$ and $\mathrm{RESIM}_{\mathrm{GW}}$ can be considered a generalization of MESIM and RESIM (Cerón-Rojas *et al*. 2008 a, b) for the case where individuals are selected based on their performance for traits measured in several environments where additional random variables are represented by molecular markers. Similarly, LW and $\mathrm{KN}_{\mathrm{GW}}$ are generalizations of the genomewide selection index of Lange and Whittaker (2001) and the Kempthorne and Nordskog (1959) restrictive selection index, respectively. The sampling properties of $\mathrm{MESIM}_{\mathrm{GW}}$, MESIM, and $\mathrm{RESIM}_{\mathrm{GW}}$, and their selection responses are known, and their estimators showed desirable statistical properties such as consistency and asymptotic unbiasedness. MESIM maximizes the selection response by maximizing the square of the correlation between Y_s and Z_s, $\rho^2_{Y_s Z_s}$, which is the same as maximizing $(\boldsymbol{\theta}_{gs}\mathbf{W}\boldsymbol{\beta}_{ps})^2$. This basic idea, used for developing a molecular selection index based on eigenanalysis (Cerón-Rojas *et al*. 2008b), is valid for $\mathrm{MESIM}_{\mathrm{GW}}$ when molecular markers are incorporated as additional traits.

REFERENCES

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*. 3rd ed. John Wiley and Sons, New Jersey.

Bailey, N.T.J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford University Press, London.

Beavis, W.D. and Keim, P. (1996). Identification of quantitative trait loci that are affected by environment, pp. 123-149 In : *Genotype by Environment Interaction*. Kang, M.S., and H.G. Gauch (eds.), CRC Press, Boca Raton, New York, USA.

Bernardo, R. (2002). *Breeding for Quantitative Traits in Plants*. Stemma Press, Woodbury, Minnesota, USA.

Bernardo, R. and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.*, **47**, 1082-1090.

Bulmer, M.G. (1980). *The Mathematical Theory of Quantitative Genetics. Lectures in Biomathematics*. University of Oxford, Clarendon Press.

Cerón-Rojas, J.J., Crossa, J., Sahagún-Castellanos, J., Castillo-González, F. and Santacruz-Varela, A. (2006). A selection index method based on eigenanalysis. *Crop Sci.*, **46**, 1711-1721.

Cerón-Rojas, J.J., Sahagún-Castellanos, J., Castillo-González, F., Santacruz-Varela, A., and Crossa, J., (2008a). A restricted selection index method based on eigenanalysis. *J. Agric. Bio. Environ. Statist*. (in press).

Cerón-Rojas, J.J., Castillo-González, F., Sahagún-Castellanos, J., Santacruz-Varela, A., Benítez-Riquelme, I. and Crossa, J. (2008b). A molecular selection index method based on eigenanalysis. *Genetics*, **180**, 547-557.

Crossa, J., Burgueño, J., Dreisigacker, S., Vargas, M., Herrera-Foessel, S.A., Lillemo, M., Singh, R.P., Trethowan, R., Warburton, M., Franco, J., Reynolds, M., Crouch, J.H. and Ortiz, R. (2007). Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. *Genetics*, DOI: 10.1534/genetics. 107.078659.

Crossa, J., Vargas, M., Van Eeuwijk, F.A., Jiang, C., Edmeades, G.O. and Hoisington, D. (1999). Interpreting genotype × environment interaction in tropical maize using linked molecular markers and environmental covariables. *Theo. Appl. Genet*., **99**, 611-625.

Falconer, D.S. and Mackay, T.F.C. (1997). *Introduction to Quantitative Genetics*. Longman, New York, 464.

Gianola, D., Pérez-Enciso, M. and Toro, M.A. (2003). On marker-assisted prediction of genetic value: Beyond the ridge. *Genetics*, **163**, 347-365.

Gianola, D., Fernando, R.L. and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, **173**, 1761-1776.

Gianola, D., and Van Kaam, J.B.C.H.M. (2008). Reproducing Kernel Hilbert spaces methods for genomic assisted prediction of quantitative traits. *Genetics*, **178**, 2289-2303.

Gimelfarb, A. and Lande, R. (1994). Simulation of marker-assisted selection in hybrid populations. *Genet. Res.*, **63**, 39-47.

Gimelfarb, A. and Lande, R. (1995). Marker-assisted selection and marker-QTL associations in hybrid populations. *Theo. Appl. Genet*., **91**, 522-528.

Harris, D.L. (1964). Expected and predicted progress from index selection involving estimates of population parameters. *Biometrics*, **20**, 46-72.

Hayes, J.F. and Hill, W.G. (1980). A reparameterization of a genetic selection index to locate its sampling properties. *Biometrics*, **36**, 237-248.

Hazel, L.N. (1943). The genetic basis for constructing a selection index, pp. 316-330 in *Papers on Quantitative Genetics and Related Topics*. Department of Genetics, North Carolina State College, Raleigh, North Carolina, USA.

Kempthorne, O. and Nordskog, A.W. (1959). Restricted selection indices. *Biometrics*, **15**, 10-19.

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

Lange, C. and Whittaker, J.C. (2001). On the prediction of genetic values in marker-assisted selection. *Genetics*, **159**, 1375-1381.

Li, Z. (1998). Molecular analysis of epistasis, pp. 119-130 in *Molecular Dissection of Complex Traits*. Paterson, A.H. (Ed.), CRC Press, Boca Raton, New York, USA

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1982). *Multivariate Analysis*. Academic Press Inc, New York, USA.

Rao, C.R. (2002). *Linear Statistical Inferences and its Applications*. 2nd Edition, John-Wiley and Sons, Inc.

Smith, H.F. (1936). A discriminant function for plant selection, pp. 466- 476 in *Papers on Quantitative Genetics and Related Topics*. Department of Genetics, North Carolina State College, Raleigh, North Carolina, USA.

Smith, O.S., Hallauer, A.R. and Russell, W.A. (1981). Use of index selection in recurrent selection programs in maize. *Euphytica*, **30**, 611-618.

Wang, J., van Ginkel, M., Podlich, D., Ye, G., Trethowan, R., Pfeiffer, W., Delacy, I.H., Cooper, M. and Rajaram, S. (2003). Comparison of two breeding strategies by computer simulation. *Crop Sci*., **43**, 1764-1773.

Wang, J., van Ginkel, M., Trethowan, R., Ye, G., Delacy, I., Podlich, D. and Cooper, M. (2004). Simulating the effects of dominance and epistasis on selection response in the CIMMYT wheat breeding program using QuCim. *Crop Sci.,* **44**, 2006-2018.

Zhang, W. and Smith, C. (1992). Computer simulation of marker-assisted selection utilizing linkage disequilibrium. *Theo. Appl. Genet.*, **83**, 813-820.

Zhang, W. and Smith, C. (1993). Simulation of marker-assisted selection utilizing linkage disequilibrium: the effects of several additional factors. *Theo. Appl. Genet.*, **86**, 492-496.

APPENDIX

Derivation of Matrices M and $\mathbf{G}_M$

The structure of the variance-covariance matrices between molecular markers (MM), **M**, and the variance-covariance between the code values of the m^{th} MM and the genotypic value of the t^{th} trait that is influenced by the additive effect of the q^{th} QTL ($\mathbf{G}_M$) requires information on the recombination frequency between MM and QTLs. Suppose that the order of any three *loci* A, B, and C is ABC (the *loci* can be MM or QTLs), and that the recombination frequency is r_1 between A and B, r_2 between B and C, and r between A and C. There are two ways of obtaining the recombination frequency between A and C: (1) recombination between A and B and no recombination between B and C; and (2) recombination between B and C and no recombination between A and B. Without interference, the recombination frequency r can be written as

$$r = r_1(1 - r_2) + (1 - r_1)r_2 = r_1 + r_2 - 2r_1r_2 \qquad \text{(A.1)}$$

also known as the TROW formula (Bailey 1961).

Consider a doubled haploid progeny having markers A and B with alleles A_1 and A_2, B_1 and B_2, respectively; suppose that a QTL with alleles Q_1 and Q_2 is located between the two markers. Denote α as the additive value of genotype Q_1Q_1 and $-\alpha$ as the additive value of genotype Q_2Q_2. Let 1 be the code value of genotypes A_1A_1, Q_1Q_1, and B_1B_1, and –1 the code value of genotypes A_2A_2, Q_2Q_2, and B_2B_2. Complete information on the doubled haploid progeny is shown in Table A.1.

From Table A.1 it is possible to obtain the expected value [$E(X_i)$], variances [$Var(X_i)$], covariances [$Cov(X_i, X_j)$], and correlations [$Corr(X_i, X_j)$], i, j = 1, 2, 3, of the code values of the MM (X_1, and X_3) and the QTL (X_2). The expected value of X_1 and the variance of X_1 are

$$\begin{aligned} E(X_1) = \frac{1}{2}\, &[(1 - r_1)(1 - r_2) + (1 - r_1)r_2 \\ &+ r_1r_2 - r_1(1 - r_2) + r_1(1 - r_2) \\ &- (1 - r_1)r_2 - r_1r_2 - (1 - r_1)(1 - r_2)] = 0 \\ V(X_1) = (1 - r_1)&(1 - r_2) + (1 - r_1)r_2 \\ &+ r_1(1 - r_2) + r_1r_2 = 1 \end{aligned}$$

The expected values and variances of X_2 and X_3, calculated in a similar manner, are, respectively, $E(X_2) = E(X_3) = 0$, and $Var(X_2) = Var(X_3) = 1$. The covariance between X_1 and X_2, X_1 and X_3, X_2 and X_3 are, respectively, $Cov(X_1, X_2) = (1 - r_1)(1 - r_2) + (1 - r_1)r_2 - r_1(1 - r_2) - r_1r_2 = 1 - 2r_1$, $Cov(X_2, X_3) = 1 - 2r_2$, and $Cov(X_1, X_3) = 1 - 2r$. The correlation between X_1 and X_2 is $Corr(X_1, X_2) = \dfrac{Cov(X_1, X_2)}{\sqrt{Var(X_1)Var(X_2)}} = 1 - 2r_1$. The other correlations can be calculated in a similar manner. Note that if the additive QTL genotype values (α_{QTL})

Table A.1. Genotypes of molecular markers (MM) and one QTL, expected genotypic frequencies, code values of the genotypes of molecular marker A (X_1), the QTL (X_2) and molecular marker B (X_3), and the additive value of the genotype of the QTL (α_{QTL}) in a double haploid population

Genotypes of the MM and the QTL	Expected frequencies	Code values of the genotypes			Additive value of the QTL genotype (α_{QTL})
		X_1	X_2	X_3	
$A_1A_1Q_1Q_1B_1B_1$	$(1 - r_1)(1 - r_2)/2$	1	1	1	α
$A_1A_1Q_1Q_1B_2B_2$	$(1 - r_1)r_2/2$	1	1	–1	α
$A_1A_1Q_2Q_2B_1B_1$	$r_1r_2/2$	1	–1	1	$-\alpha$
$A_2A_2Q_1Q_1B_1B_1$	$r_1\ (1 - r_2)/2$	–1	1	1	α
$A_1A_1Q_2Q_2B_2B_2$	$r_1\ (1 - r_2)/2$	1	–1	–1	$-\alpha$
$A_2A_2Q_2Q_2B_1B_1$	$(1 - r_1)r_2/2$	–1	–1	1	$-\alpha$
$A_2A_2Q_1Q_1B_2B_2$	$r_1r_2/2$	–1	1	–1	α
$A_2A_2Q_2Q_2B_2B_2$	$(1 - r_1)(1 - r_2)/2$	–1	–1	–1	$-\alpha$

are used instead of the code values of the genotypes of the QTL(X_2), then the covariance between the code value of the first marker (X_1) and α_{QTL} is $Cov(X_1, \alpha_{\text{QTL}}) = (1 - r_1)(1 - r_2)\alpha + (1 - r_1)r_2\ \alpha - r_1(1 - r_2)\alpha - r_1r_2\alpha = (1 - 2r_1)\alpha$.

In addition, it is worth noting that the MM are variables not affected by the environment or the experimental error; thus the covariance between the MM's phenotypic values and code values is equal to the covariance of the genotypic values with the code values of the MM. Thus, suppose that $p_{eitm} = g_{eitm} + E_e + GE_{ie} + \varepsilon_{eit}$, where m^{th} denotes the MM affecting the i^{th} genotypic value of the t^{th} trait in the e^{th} environment ($m = 1, 2, \ldots, N_m$; $t = 1, 2, \ldots, N_t$; $i = 1, 2, \ldots, n$), and that X_m is the m^{th} random variable that denotes the code values of the m^{th} MM; if $E(E_e) = E(GE_{ie}) = E(\varepsilon_{eit}) = 0$, then $Cov(p_{eitm}, X_m) = Cov(g_{itm}, X_m)$. This allows writing matrices $\mathbf{M}$, $\mathbf{G}_M$, $\mathbf{\Gamma}$, and $\mathbf{\Sigma}$ as they were written in the text.