

Efficiency of a Constrained Linear Genomic Selection Index To Predict the Net Genetic Merit in Plants

J. Jesus Cerón-Rojas and Jose Crossa¹

Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600, Mexico City, Mexico

ORCID IDs: 0000-0003-2885-6831 (J.J.C.-R.); 0000-0001-9429-5855 (J.C.)

ABSTRACT The constrained linear genomic selection index (CLGSI) is a linear combination of genomic estimated breeding values useful for predicting the net genetic merit, which in turn is a linear combination of true unobservable breeding values of the traits weighted by their respective economic values. The CLGSI is the most general genomic index and allows imposing constraints on the expected genetic gain per trait to make some traits change their mean values based on a predetermined level, while the rest of them remain without restrictions. In addition, it includes the unconstrained linear genomic index as a particular case. Using two real datasets and simulated data for seven selection cycles, we compared the theoretical results of the CLGSI with the theoretical results of the constrained linear phenotypic selection index (CLPSI). The criteria used to compare CLGSI vs. CLPSI efficiency were the estimated expected genetic gain per trait values, the selection response, and the interval between selection cycles. The results indicated that because the interval between selection cycles is shorter for the CLGSI than for the CLPSI, CLGSI is more efficient than CLPSI per unit of time, but its efficiency could be lower per selection cycle. Thus, CLGSI is a good option for performing genomic selection when there are genotyped candidates for selection.

KEYWORDS

Genomic estimated breeding value
Selection response
Expected genetic gain per trait
Molecular marker effects
GenPred
Shared Data Resources
Genomic Prediction

The unconstrained linear genomic selection index (LGSI) is a linear combination of genomic estimated breeding values (GEBVs) and was originally proposed by Togashi *et al.* (2011); however, Cerón-Rojas *et al.* (2015) developed the LGSI theory completely and applied the LGSI theoretical results to real and simulated data. The LGSI is useful for predicting the net genetic merit, which in turn is a linear combination of true unobservable breeding values of the traits weighted by their respective economic values (Hazel 1943). In LGSI, all marker effects of

the genotyped individuals in the training population are estimated using marker and phenotypic data; these estimated effects are then used in subsequent selection cycles to obtain GEBVs that are predictors of the individual breeding values in the testing population for which there is only marker information about the candidates for selection. In LGSI, the GEBVs can be obtained by multiplying the genomic best linear unbiased predictor (GBLUP) of the estimated marker effects in the training population (VanRaden 2008) by the coded marker values obtained in the testing population in each selection cycle. Applying LGSI in plant or animal breeding requires genotyping the candidates for selection to obtain marker effects and GEBVs, and then predicting and ranking the net genetic merit of the candidates for selection.

The LGSI was developed in the genomic selection (GS) context in which animals and plants are selected based on the GEBV of the candidates for selection. Meuwissen *et al.* (2001) developed the GS theory and showed that it increases the accuracy of predicting the breeding values of the candidates for selection, and reduces the intervals between selection cycles and the costs of the breeding programs. Meuwissen *et al.* (2001) suggested estimating all marker effects jointly

Copyright © 2019 Cerón, Crossa

doi: <https://doi.org/10.1534/g3.119.400677>

Manuscript received August 28, 2019; accepted for publication September 26, 2019; published Early Online September 30, 2019.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

¹Corresponding author: Biometrics and Statistics Unit, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600, Mexico City, Mexico. E-mail: j.crossa@cgiar.org

using linear mixed models and Bayesian methods to increase the accuracy of the predicted breeding values because many genes affect the quantitative traits, which are of most concern to plant and animal breeders. Isik *et al.* (2017), however, have indicated that GS has not changed the fundamental basis of breeding methods, since in GS the individuals are ranked based on their estimated breeding values, which has been the selection method used by animal and plant breeders for a long time. Nevertheless, because GS decreases the generation interval, it leads to a much higher genetic gain per year. For example, in the plant breeding context, a four-year breeding cycle, which includes three years of field testing, can be reduced to only four months, *i.e.*, the time required to grow and cross plants (Lorenz *et al.* 2011). Börner and Reinsch (2012) have indicated that GS could replace traditional progeny testing when maximizing the genetic gain per year, as long as the accuracy of GEBV is higher than or equal to 0.45.

One of the main problems associated with the LGSI theory is that the values of its expected genetic gain per trait (or multi-trait selection response) can increase or decrease in a positive or negative direction without control. In the phenotypic selection context, Kempthorne and Nordskog (1959) developed a restricted linear selection index that allows imposing restrictions equal to zero on the expected genetic gain of some traits. Other authors (Mallard 1972; Harville 1975; Tallis 1985) extended the Kempthorne and Nordskog (1959) approach and developed a constrained linear phenotypic selection index (CLPSI) that attempts to make some traits change their expected genetic gain values based on a predetermined level, while the rest of the traits remain without restrictions. Itoh and Yamada (1987) showed that the Mallard (1972), Harville (1975) and Tallis (1985) indices give the same results. The CLPSI is the most general index and includes the unconstrained and restricted phenotypic indices as particular cases.

Céron-Rojas and Crossa (2018, Chapter 3) developed a constrained linear genomic selection index (CLGSI); however, these authors did not evaluate this index completely. For example, they did not give results associated with the CLGSI expected genetic gain per trait, which is the main parameter of this index because the breeder imposes constraints on it to make some traits change their expected genetic gain values based on a predetermined level, while the rest of the traits remain without restrictions.

This study had two main objectives: first, to apply the CLGSI theoretical results to two real datasets and to seven simulated datasets using only GEBVs for selecting non-phenotyped candidates for selection, and second, to compare the relative efficiency of CLGSI and CLPSI using real and simulated datasets in a single-stage context. The criteria we used to compare CLGSI vs. CLPSI efficiency were the estimated expected genetic gain per trait values, the selection response, and the interval between selection cycles. One additional criterion was that the CLGSI and CLPSI proportional constant value should be positive (Céron-Rojas and Crossa 2018, Chapters 3 and 6 for details). The results indicated that because the interval between selection cycles was shorter for CLGSI than for CLPSI, CLGSI is more efficient than CLPSI per unit of time, but its efficiency could be lower per selection cycle.

We applied the CLGSI theory assuming that the GEBV values have multivariate normal distribution and that the CLGSI (CLPSI) and the net genetic merit have joint bivariate normal distribution. Under this last assumption, the regression of the net genetic merit on any linear function of the phenotypic or GEBV values is linear (Kempthorne and Nordskog 1959).

MATERIAL AND METHODS

Objectives of the constrained linear selection index

Céron-Rojas *et al.* (2016) and Céron-Rojas and Crossa (2018, Chapter 3) described the CLPSI theory; thus, in this work, we shall describe only the CLGSI theory. Let μ_i be the population mean of the i^{th} trait before selection. One of the main CLGSI objectives is to change μ_i to $\mu_i + d_i$, where d_i is the i^{th} ($i = 1, 2, \dots, r$; $r =$ number of constrained traits) trait

constraint or predetermined proportional gain imposed by the breeder on the CLGSI expected genetic gain per trait. Additional CLGSI objectives are to maximize the selection response; to predict the net genetic merit ($H = \mathbf{w}'\mathbf{g}$, where $\mathbf{w}' = [w_1 \ w_2 \ \dots \ w_t]$ and $\mathbf{g}' = [g_1 \ g_2 \ \dots \ g_t]$ are $1 \times t$ ($t =$ number of traits) vectors of economic weights and true unobservable breeding values, respectively); to select individuals with the highest H values as parents of the next generation; and to provide the breeder with an objective rule for evaluating and selecting several traits simultaneously.

The constrained linear genomic selection index

Let $\mathbf{z}' = [z_1 \ z_2 \ \dots \ z_t]$ be a vector of genomic breeding values for t traits (Appendix 1, Equations A1 to A4 for details). The CLGSI used to predict the individual net genetic merit of a candidate for selection is

$$I = \beta_1 z_1 + \beta_2 z_2 + \dots + \beta_t z_t = \boldsymbol{\beta}' \mathbf{z}, \quad (1)$$

where $\boldsymbol{\beta}' = [\beta_1 \ \beta_2 \ \dots \ \beta_t]$ is the CLGSI vector of coefficients and z_i ($i = 1, 2, \dots, t$) is the genomic value associated with trait i^{th} . Céron-Rojas *et al.* (2015) described Equation (1) in the unconstrained LGSI.

The CLGSI selection response

The CLGSI selection response (R) can be written as

$$R = \frac{1}{L} k \sigma_H \rho_{HI}, \quad (2)$$

where k is the selection intensity, L denotes the interval between selection cycles, $\sigma_H = \sqrt{\mathbf{w}'\mathbf{C}\mathbf{w}}$ is the standard deviation of $H = \mathbf{w}'\mathbf{g}$, $\text{Var}(\mathbf{g}) = \mathbf{C}$ is the covariance matrix of \mathbf{g} , and $\rho_{HI} = \frac{\mathbf{w}'\boldsymbol{\Gamma}\boldsymbol{\beta}}{\sqrt{\mathbf{w}'\mathbf{C}\mathbf{w}}\sqrt{\boldsymbol{\beta}'\boldsymbol{\Gamma}\boldsymbol{\beta}}}$ is the correlation between $H = \mathbf{w}'\mathbf{g}$ and $I = \boldsymbol{\beta}'\mathbf{z}$, where $\mathbf{w}'\boldsymbol{\Gamma}\boldsymbol{\beta} = \sigma_{HI}$ is the covariance between $H = \mathbf{w}'\mathbf{g}$ and $I = \boldsymbol{\beta}'\mathbf{z}$, $\boldsymbol{\Gamma} = \text{Var}(\mathbf{z})$ is the covariance matrix of genomic breeding values (Appendix 1, Equation A2 for details), and $\sigma_I = \sqrt{\boldsymbol{\beta}'\boldsymbol{\Gamma}\boldsymbol{\beta}}$ is the standard deviation of $I = \boldsymbol{\beta}'\mathbf{z}$. Equation (1) indicates that the genetic gain that can be achieved in R by selecting for several traits simultaneously within a population of plants, is the product of k , σ_H and ρ_{HI} (Kempthorne and Nordskog 1959). Thus, to increase selection progress, ρ_{HI} should be as large as possible. Céron-Rojas and Crossa (2018, Chapter 3) described the selection response of the CLPSI, which is very similar to Equation (2).

The CLGSI expected genetic gain per trait

The CLGSI expected genetic gain per trait (\mathbf{E} , or multi-trait selection response) is the covariance between the genomic breeding value vector \mathbf{z}' and the index, $I = \boldsymbol{\beta}'\mathbf{z}$ (Equation 1), weighted by the standard deviation of I , $\sigma_I = \sqrt{\boldsymbol{\beta}'\boldsymbol{\Gamma}\boldsymbol{\beta}}$, multiplied by the selection intensity (k) and divided by the interval between selection cycles (L), *i.e.*,

$$\mathbf{E} = \frac{k}{L} \frac{\boldsymbol{\Gamma}\boldsymbol{\beta}}{\sigma_I}. \quad (3)$$

Equation (3) is a $t \times 1$ vector, *i.e.*, $\mathbf{E}' = [E_1 \ E_2 \ \dots \ E_t]$, where $E_i = \frac{k}{L\sigma_I} \text{Cov}(I, z_i) = \frac{k}{L\sigma_I} [\beta_1 \sigma_{1i} + \beta_2 \sigma_{2i} + \dots + \beta_j \sigma_{ji} + \dots + \beta_t \sigma_{ti}]$ ($i = 1, 2, \dots, t$) is the expected genetic gain of trait i^{th} , z_i is the i^{th} genomic breeding value, $I = \sum_{j=1}^t \beta_j z_j$ is the index value (Equation 1) for one individual, $\text{Cov}(I, z_i)$ is the covariance between I and z_i , and σ_{ji} is the covariance between z_i and the j^{th} ($j = 1, 2, \dots, t$) index genomic breeding value. We defined all the other terms of Equation (3) in Equation (2).

The CLGSI vector of coefficients

Maximizing ρ_{HI} (Equation 2) is equivalent to minimizing the mean squared difference between the net genetic merit $H = \mathbf{w}'\mathbf{g}$ and the

index $I = \beta'z$, i.e., $E[(H-I)^2]$, with respect to the vector of coefficients β under the restriction that Equation (3) values are equal to the d_i values ($i = 1, 2, \dots, r$) imposed by the breeder. Details of the process used to obtain the CLGSI vector of coefficients (β) are given in **Appendix 2** (Equations A6 and A7). Here we present only the main result. The CLGSI vector of coefficients is

$$\beta = Kw, \quad (4a)$$

where $w' = [w_1 \ w_2 \ \dots \ w_t]$ is a vector of economic weights, $K = [I_t - Q]$, $Q = UD(D'U'ΓUD)^{-1}D'U'Γ$, I_t is an identity matrix of size $t \times t$, D is a Mallard (1972) matrix described in **Appendix 2**, and U is the Kempthorne and Nordskog (1959) matrix described in **Appendix 3**. When $D = U$ ($d_i = 0$ for all traits), matrix K can be written as $K_G = [I_t - Q_G]$, where $Q_G = U(U'ΓU)^{-1}U'Γ$. In this last case, the CLGSI is a null restricted LGSI similar to the Kempthorne and Nordskog (1959) restricted index. When $D = U$ and U' is a null matrix, $\beta = w$, the vector of economic weights or the unconstrained LGSI vector of coefficients (Ceron-Rojas *et al.* 2015). Thus, the CLGSI is more general than the LGSI and includes the null restricted and the unrestricted LGSI as particular cases. The vector of coefficients β should maximize the selection response (R , Equation 2) and make Equation (3) values similar to d_i values.

It is possible to show (Itoh and Yamada 1987; Céron-Rojas and Crossa 2018, Chapter 3) that another way of writing Equation (4a) is

$$\beta = K_G w + \theta U(U'ΓU)^{-1}d, \quad (4b)$$

where $\theta = \frac{d'(U'ΓU)^{-1}U'Γw}{d'(U'ΓU)^{-1}d}$ is the proportionality constant and $d' = [d_1 \ d_2 \ \dots \ d_r]$ is the vector of the constraints imposed by the breeder. According to Itoh and Yamada (1987), θ should be higher than zero ($\theta > 0$) because when $\theta < 0$, the CLGSI moves the population means in the opposite direction to the predetermined desired direction. In addition, when $\theta = 0$, we would have the CLGSI with null constraints similar to the Kempthorne and Nordskog (1959) restricted index. Thus, θ is a good criterion for determining when the CLGSI moves the population means in the desired direction, which will occur when $\theta > 0$. While Equation (4a) is associated with the Mallard (1972) constrained index, Equation (4b) is associated with the Tallis (1985) constrained index; however, Equations (4a) and (4b) express the same result in a different mathematical way.

In **Appendix 1** (Equation A5), we give a method for estimating matrix Γ , and in **Appendix 2** (Equation A8), we explain how to estimate β .

Maximized CLGSI selection response and optimized expected genetic gain per trait

The maximized CLGSI selection response is

$$R = k\sqrt{\beta'\Gamma\beta}, \quad (5)$$

while the optimized CLGSI expected genetic gain per trait is the same as Equation (3). In Equation (5) we omitted L to simplify notation. Whereas in Equation (2) the selection response can take any value, in Equation (5), R gives the maximum value of Equation (2). In **Appendix 2** (Equations A9 and A10), we indicate how to estimate R and E .

Obtaining the genomic estimated breeding values (GEBV)

Several authors (VanRaden 2008; Ceron-Rojas *et al.* 2015; Isik *et al.* 2017, Chapter 11; Céron-Rojas and Crossa 2018, Chapter 5) have given detailed descriptions of how to obtain trait genomic breeding values (GEBV) which are predictors of trait unobservable breeding values. In CLGSI, we fitted phenotypic and marker data from the training

population in a statistical model to estimate all available marker effects; these estimates were then used to obtain GEBV that are predictors of the individual traits' true genomic breeding values in a testing population for which there is only marker information. We obtained the GEBV in the non-phenotyped testing population by multiplying the estimated marker effects obtained in the training population by the coded marker values obtained in the testing population at each selection cycle.

Criteria for comparing CLGSI efficiency vs. CLPSI efficiency

The criteria used to compare CLGSI vs. CLPSI efficiency when making genomic and phenotypic selection were as follows. First, the estimated theta value (θ) should be positive ($\theta > 0$). Second, the estimated expected genetic gain values should be close to the constraints (d_i) imposed by the breeder. Third, the estimated selection response value should be equivalent to the true selection response value. Fourth, the size of the interval between selection cycles (L) should be short to increase the selection gain per year.

In this work, we do not consider the problem associated with the cost of obtaining measures of each phenotypic trait and the process of genotyping individual candidates for selection. However, some authors (Schaeffer 2006; Börner and Reinsch 2012) have considered this problem in the animal breeding context.

MATERIALS

Real data

We used two real maize (*Zea mays* L.) F2 populations: "JMpop1 DTMA Mexico optimum environment" and "6x1020 WEMA Africa optimum environment" (hereafter we shall refer to the first and second maize F2 populations as dataset 1 and dataset 2, respectively). The training population (C0) of each dataset contains genotypic data and four phenotypic traits: grain yield (GY, t/ha), plant height (PHT, cm), ear height (EHT, cm), and anthesis days (AD, d). In addition, each dataset has three sets of individuals from the training population (C0) and two testing populations (C1 and C2). We present the number of individuals and molecular markers in each population in Table 1. Assuming that the breeding objective was to increase GY while decreasing PHT, EHT, and AD, the vector of economic weights in C0, C1, and C2 for GY, PHT, EHT, and AD was $w' = [5 \ -0.3 \ -0.3 \ -1]$ for both indices and the two datasets. For illustration purposes only, in this work we used three selection cycles (C0, C1 and C2) for both datasets to illustrate the theoretical results and the efficiency of CLGSI and CLPSI.

For illustration purposes only, to select traits GY, PHT, EHT and AD, we imposed two sets of restrictions. First we restricted traits GY and PHT with vector $d' = [0.5 \ -1.0]$ and matrices $D' = [-1.0 \ -0.5]$ and $U' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$ (**Appendices 2 and 3** for details), and later, we restricted traits GY, PHT and EHT with vector

$$d' = [0.5 \ -1.0 \ -0.5] \text{ and matrices } U' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \text{ and}$$

$D' = \begin{bmatrix} -0.5 & 0 & -0.5 \\ 0 & -0.5 & 1.0 \end{bmatrix}$ for both datasets and for the two indices (CLGSI and CLPSI). The total proportion (p) of retained value for these datasets was $p = 0.10$ ($k = 1.755$) for both indices and the two datasets with the two sets of constraints.

Simulated datasets

The datasets were simulated by Ceron-Rojas *et al.* (2015) with QU-GENE software (Podlich and Cooper 1998) using 2500 molecular markers and 315 quantitative trait loci (QTL) for eight phenotypic selection cycles

■ **Table 1** Two real maize (*Zea mays* L.) F₂ populations and the number of genotypes (*g*) and molecular markers (*m*) used in three selection cycles (cycles 0, 1 and 2)

Cycle	Real datasets			
	Dataset 1		Dataset 2	
	<i>g</i>	<i>m</i>	<i>g</i>	<i>m</i>
0	247	195	181	205
1	320	195	274	205
2	303	195	274	205

(C0 to C7), each with four traits (*T*₁, *T*₂, *T*₃ and *T*₄), 500 genotypes and 4 replicates for each genotype. The authors distributed the markers uniformly across 10 chromosomes and the QTL randomly across the 10 chromosomes to simulate maize (*Zea mays* L.) populations. A different number of QTL affected each of the four traits: 300, 100, 60, and 40, respectively. The common QTL affecting the traits generated genotypic correlations of -0.5, 0.4, 0.3, -0.3, -0.2, and 0.1 between *T*₁ and *T*₂, *T*₁ and *T*₃, *T*₁ and *T*₄, *T*₂ and *T*₃, *T*₂ and *T*₄, *T*₃ and *T*₄, respectively. The

economic weights for *T*₁, *T*₂, *T*₃ and *T*₄ were 1, -1, 1 and 1, respectively. Additional details of the simulated data can be seen in Cerón-Rojas *et al.* (2015).

We used seven phenotypic and genomic selection cycles (C1 to C7) with *p* = 0.10 (*k* = 1.755) in each cycle. We selected all four traits in each selection cycle. For illustration purposes only, to select traits *T*₁, *T*₂, *T*₃ and *T*₄, we imposed two sets of restrictions. First we restricted traits *T*₁ and *T*₂ with vector **d'** = [5 -2] and matrices

$$\mathbf{U}' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \text{ and } \mathbf{D}' = [-2 \quad -5], \text{ and later, we restricted}$$

traits *T*₁, *T*₂ and *T*₃ with vector **d'** = [5 -2 3] and matrices

$$\mathbf{U}' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \text{ and } \mathbf{D}' = \begin{bmatrix} 3 & 0 & -5 \\ 0 & 3 & 2 \end{bmatrix} \text{ for both datasets and}$$

both indices.

Real and simulated data availability

The real and simulated datasets are available in the *Application of a Genomics Selection Index to Real and Simulated Data* repository, at

■ **Table 2** Two real datasets for estimated CLGSI^a parameters constructed with four traits (GY, PHT, EHT and AD) for two and three constraints in three selection cycles

Cycle	Dataset 1 with two constraints					
	Theta	Estimated expected genetic gain per trait				Estimated maximum Response
		GY	PHT	EHT	AD	
0	2.32	0.76	-1.51	-4.05	0.71	4.73
1	1.70	0.65	-1.30	-3.47	0.67	4.02
2	1.57	0.61	-1.22	-3.63	0.56	3.96
Mean1 ^b	1.86	0.67	-1.35	-3.72	0.65	4.24
Mean2 ^c	—	0.45	-0.90	-2.48	0.43	2.82

Cycle	Dataset 1 with three constraints					
	Theta	Estimated expected genetic gain per trait				Estimated maximum Response
		GY	PHT	EHT	AD	
0	2.74	0.84	-1.68	-0.84	0.43	5.02
1	1.99	0.72	-1.44	-0.72	0.41	4.27
2	1.89	0.70	-1.39	-0.70	0.36	4.18
Mean1 ^b	2.21	0.75	-1.50	-0.75	0.40	4.49
Mean2 ^c	—	0.50	-1.00	-0.50	0.27	2.99

Cycle	Dataset 2 with two constraints					
	Theta	Estimated expected genetic gain per trait				Estimated maximum Response
		GY	PHT	EHT	AD	
0	0.84	0.42	-0.83	-1.30	-0.36	3.08
1	0.76	0.43	-0.86	-0.61	-0.14	2.73
2	0.68	0.41	-0.82	-0.43	-0.11	2.54
Mean1 ^b	0.76	0.42	-0.84	-0.78	-0.20	2.79
Mean2 ^c	—	0.28	-0.56	-0.52	-0.13	1.86

Cycle	Dataset 2 with three constraints					
	Theta	Estimated expected genetic gain per trait				Estimated maximum Response
		GY	PHT	EHT	AD	
0	0.94	0.45	-0.89	-0.45	-0.36	3.25
1	0.86	0.44	-0.88	-0.44	-0.12	2.99
2	0.75	0.41	-0.83	-0.41	-0.10	2.79
Mean1 ^b	0.85	0.43	-0.87	-0.43	-0.19	3.01
Mean2 ^c	—	0.29	-0.58	-0.29	-0.13	2.01

^aConstrained Linear Genomic Selection Index.

^bMean1 is the average of the three selection cycles.

^cMean2 = Mean1/1.5, where 1.5 is the interval between selection cycles and denotes the average of the genetic gain per year.

■ **Table 3** Two real datasets for estimated LGSI^a parameters constructed with four traits (GY, PHT, EHT and AD) without constraints for three selection cycles

Cycle	Dataset 1					Estimated maximum Response
	Estimated expected genetic gain per trait				Estimated maximum Response	
	GY	PHT	EHT	AD		
0	0.23	-6.05	-6.45	-0.15	5.05	
1	0.21	-5.15	-5.47	-0.07	4.29	
2	0.18	-4.94	-5.60	-0.15	4.21	
Mean1 ^b	0.21	-5.38	-5.84	-0.12	4.52	
Mean2 ^c	0.14	-3.59	-3.89	-0.08	3.01	

Cycle	Dataset 2					Estimated maximum Response
	Estimated expected genetic gain per trait				Estimated maximum Response	
	GY	PHT	EHT	AD		
0	0.33	-2.65	-2.47	-0.36	3.57	
1	0.35	-1.73	-1.44	-0.17	2.85	
2	0.29	-1.81	-1.61	-0.19	2.67	
Mean1 ^b	0.32	-2.06	-1.84	-0.24	3.03	
Mean2 ^c	0.22	-1.37	-1.23	-0.16	2.02	

^aUnconstrained Linear Genomic Selection Index.

^bMean1 is the average of the three selection cycles.

^cMean2 = Mean1/1.5, where 1.5 is the interval between selection cycles and denotes the average of the genetic gain per year.

<http://hdl.handle.net/11529/10199>. The two real datasets used in this work are the folder named "File Real_Data_Sets_GSP" that contains four folders called "DATA_SET-3, 4, 5 and 6". Each of the four folders in turn contains

four Excel data files. The four Excel data files within the folder DATA_SET-3 are as follows: DATA_SET-3_Markers_Cycle-0, 1, 2, and DATA_SET-3_Phenotypic_Cycle-0. The first three Excel files contain the marker

■ **Table 4** Simulated data for estimated CLPSI^a and CLGSI^b parameters constructed with four traits (T1, T2, T3 and T4), two constraints and true maximum responses in seven selection cycles

Cycle	Theta	CLPSI				Estimated maximum Response	True maximum Response
		Estimated expected genetic gains per trait					
		T1	T2	T3	T4		
1	8.37	7.90	-3.16	3.55	1.70	16.31	19.51
2	7.53	7.48	-2.99	3.23	1.79	15.50	17.56
3	6.34	6.93	-2.77	2.74	1.65	14.08	16.49
4	7.19	7.75	-3.10	2.17	1.25	14.28	16.30
5	6.21	7.02	-2.81	2.46	1.33	13.62	15.96
6	4.92	6.30	-2.52	2.00	1.21	12.03	14.57
7	4.17	5.57	-2.23	2.58	1.17	11.54	14.65
Mean1 ^c	6.39	6.99	-2.80	2.68	1.44	13.91	16.43
Mean2 ^d	—	1.75	-0.70	0.67	0.36	3.48	4.11

Cycle	Theta	CLGSI				Estimated maximum Response	True maximum Response
		Estimated expected genetic gains per trait					
		T1	T2	T3	T4		
1	6.13	6.83	-2.73	2.79	1.47	13.82	12.65
2	5.67	6.57	-2.63	2.69	1.41	13.29	15.27
3	5.26	6.28	-2.51	2.69	1.41	12.90	15.10
4	4.36	5.79	-2.32	2.29	1.21	11.60	16.03
5	3.78	5.37	-2.15	2.15	1.17	10.84	15.17
6	3.44	5.13	-2.05	1.92	1.21	10.31	14.28
7	3.78	5.39	-2.15	2.17	1.11	10.82	15.73
Mean1 ^c	4.63	5.91	-2.36	2.38	1.28	11.94	14.89
Mean3 ^e	—	3.94	-1.58	1.59	0.86	7.96	9.92

^aConstrained Linear Phenotypic Selection Index;

^bConstrained Linear Genomic Selection Index;

^cMean1 is the average of the seven selection cycles;

^dMean2 = Mean1/ 4 for the CLPSI;

^eMean3 = Mean1/1.5 for the CLGSI, where 4 and 1.5 are the intervals between selection cycles for CLPSI and CLGSI, respectively. Means 2 and 3 denote the average of the genetic gains per year.

■ **Table 5** Simulated data for estimated CLPSI^a and CLGSI^b parameters constructed with four traits (T1, T2, T3 and T4), three constraints and true maximum responses in seven selection cycles

Cycle	Theta	CLPSI						True maximum Response
		Estimated expected genetic gains per trait				Estimated maximum Response		
		T1	T2	T3	T4			
1	7.30	7.12	-2.85	4.27	1.55	15.79	17.47	
2	6.43	6.61	-2.64	3.97	1.76	14.97	16.24	
3	5.32	6.03	-2.41	3.62	1.52	13.58	15.15	
4	4.45	5.54	-2.21	3.32	1.29	12.37	13.95	
5	4.78	5.76	-2.30	3.45	1.28	12.80	14.28	
6	3.62	4.96	-1.98	2.98	1.32	11.24	13.11	
7	3.65	5.00	-2.00	3.00	1.23	11.24	13.14	
Mean1 ^c	5.08	5.86	-2.34	3.52	1.42	13.14	14.76	
Mean2 ^d	—	1.46	-0.59	0.88	0.36	3.28	3.69	

Cycle	Theta	CLGSI						True maximum Response
		Estimated expected genetic gains per trait				Estimated maximum Response		
		T1	T2	T3	T4			
1	4.43	5.50	-2.20	3.30	1.40	12.41	11.76	
2	3.98	5.19	-2.08	3.11	1.42	11.80	13.40	
3	3.92	5.15	-2.06	3.09	1.44	11.74	13.47	
4	3.05	4.57	-1.83	2.74	1.13	10.28	14.11	
5	2.72	4.30	-1.72	2.58	1.16	9.75	13.47	
6	2.29	3.92	-1.57	2.35	1.15	8.99	13.30	
7	2.57	4.18	-1.67	2.51	1.12	9.48	14.36	
Mean1 ^c	3.28	4.69	-1.87	2.81	1.26	10.63	13.41	
Mean3 ^e	—	3.12	-1.25	1.87	0.84	7.09	8.94	

^aConstrained Linear Phenotypic Selection Index.

^bConstrained Linear Genomic Selection Index.

^cMean1 is the average of the seven selection cycles.

^dMean2 = Mean1/ 4 for the CLPSI.

^eMean3 = Mean1/1.5 for the CLGSI, where 4 and 1.5 are the intervals between selection cycles for CLPSI and CLGSI, respectively. Means 2 and 3 denote the average of the genetic gains per year.

coded values for cycles 0, 1 and 2, while the Excel file DATA_SET-3_Phenotypic_Cycle-0 contains the phenotypic information of cycle 0 (training population). The Excel data files of the other folders were described in a similar manner as for folder 3. In this work, we used datasets 3 and 6 to make selections and to estimate the theta values, the selection response and the genetic expected gains. The results are presented in Tables 2 and 3.

Folder Simulated_Data_GSI contains two folders: Data_Phenotypes_April-26-15 and Haplotypes_GSI_April-26-15. In turn, folder Data_Phenotypes_April-26-15 contains two folders: GSI_Phenotypes-05 and PSI_Phenotypes-05. Within folder GSI_Phenotypes-05, there are six Excel data files, each denoted as C2_GSI_05_Pheno, C3_GSI_05_Pheno, C4_GSI_05_Pheno, C5_GSI_05_Pheno and C6_GSI_05_Pheno, corresponding to the simulated phenotypic information for the genomic selection index for cycles 2-7. In addition, folder GSI_Phenotypes-05 contains eight Excel datasets denoted as C0_Pheno_05, C1_PSI_05_Pheno, C2_PSI_05_Pheno, C3_PSI_05_Pheno, C4_PSI_05_Pheno, C5_PSI_05_Pheno, C6_PSI_05_Pheno, and C7_PSI_05_Pheno corresponding to the phenotypic simulated information for the phenotypic selection index for cycles 0-7. File Haplotypes_GSI_April-26-15 contains the haplotypes of the markers for cycles 0-7 of GSI. We present the results of the simulated datasets in Tables 4, 5 and 6.

RESULTS

Real data

The normality assumption: Figure 1 presents the frequency distribution of the GEBVs associated with traits GY (Figure 1a) and PHT

(Figure 1b) for Dataset 1, in cycle 1. In addition, Figure 2 presents the frequency distribution of the CLGSI values for real Dataset 1 (in cycle 1) with two constraints (Figure 2a, $\mathbf{d}' = [0.5 \ -1.0]$), whereas Figure 2b presents the frequency distribution of the CLGSI values for real Dataset 2 (in cycle 2) with three constraints ($\mathbf{d}' = [0.5 \ -1.0 \ -0.5]$). Based on these results, we can assume that the GEBVs associated with traits GY and PHT, and the CLGSI values for the two set of restrictions, approach the normal distribution.

Estimated maximized CLGSI parameters: Tables 2 and 3 present the estimated CLGSI and LGSi parameters, respectively. The estimated parameters are the theta ($\hat{\theta}$) values (for CLGSI only), the expected genetic gains per trait \hat{E} (Appendix 2, Equation A10), and the selection responses \hat{R} (Appendix 2, Equation A9). In Table 2, all parameters were estimated for datasets 1 and 2 with two ($\mathbf{d}' = [0.5 \ -1.0]$) and three ($\mathbf{d}' = [0.5 \ -1.0 \ -0.5]$) constraints, whereas in Table 3, the parameters were estimated, without constraints, for three selection cycles (0, 1 and 2). The estimated theta values were all positive (Table 2) for the two real datasets, which indicates that the CLGSI moves the population means in the desired direction, as we would expect.

Estimated CLGSI expected genetic gains per trait: In Tables 2 and 3, there are two means: Mean1 and Mean2 = Mean1/L (L = interval between selection cycles) associated with the estimated CLGSI and LGSi expected genetic gain per trait and selection response for the two real maize (*Zea mays* L.) F2 populations, respectively. Mean1 is the average of the three

■ **Table 6 Simulated data for estimated LPSI^a and LGSI^b parameters constructed with four traits (T1, T2, T3 and T4) without constraints and true maximum responses for seven selection cycles**

Cycle	LPSI					Estimated maximum Response	True maximum Response
	Estimated expected genetic gain per trait						
	T1	T2	T3	T4			
1	10.42	-5.47	3.78	2.04	17.81	19.63	
2	10.11	-4.35	3.68	2.00	15.69	17.56	
3	9.91	-4.07	3.32	1.66	14.22	16.49	
4	10.94	-4.31	2.57	1.42	14.34	16.32	
5	10.60	-3.51	3.04	1.48	13.64	15.99	
6	10.02	-3.54	2.53	1.37	12.04	14.69	
7	8.77	-3.49	3.14	1.38	11.61	14.90	
Mean1 ^c	10.11	-4.11	3.15	1.62	14.19	16.51	
Mean2 ^d	2.53	-1.03	0.79	0.41	3.55	4.13	

Cycle	LGSI				Estimated maximum Response	True maximum Response
	Estimated expected genetic gain per trait					
	T1	T2	T3	T4		
1	6.60	-3.50	2.70	1.60	14.40	13.26
2	6.30	-3.40	2.60	1.50	13.91	15.28
3	6.10	-3.30	2.70	1.50	13.61	15.37
4	5.60	-3.10	2.30	1.30	12.30	16.05
5	5.20	-2.80	2.10	1.30	11.40	15.17
6	4.90	-2.60	1.90	1.30	10.61	14.49
7	5.20	-2.70	2.10	1.20	11.21	15.82
Mean1 ^c	5.70	-3.10	2.30	1.40	12.49	15.06
Mean3 ^e	3.8	-2.07	1.53	0.93	8.33	10.04

^aUnconstrained Linear Phenotypic Selection Index.

^bUnconstrained Linear Genomic Selection Index.

^cMean1 is the average of the seven selection cycles.

^dMean2 = Mean1/ 4 for the LPSI.

^eMean3 = Mean1/1.5 for the LGSI, where 4 and 1.5 are the intervals between selection cycles (L) for LPSI and LGSI, respectively. Means 2 and 3 denote the average of the genetic gains per year.

selection cycles, whereas Mean2 (Mean1/1.5, where 1.5 is the interval between selection cycles or the time required to complete one selection cycle (Beyene *et al.* 2015)) is the average of the genetic gain per year.

For dataset 1, $p = 0.10$, $k = 1.755$ and two constraints, the estimated CLGSI expected genetic gains per trait for cycles 0, 1 and 2 were $\hat{E}'_0 = [0.76 \ -1.51 \ -4.05 \ 0.71]$, $\hat{E}'_1 = [0.65 \ -1.30 \ -3.47 \ 0.67]$ and $\hat{E}'_2 = [0.61 \ -1.22 \ -3.63 \ 0.56]$, respectively. Each \hat{E}'_q ($q = 0, 1, 2$) value was associated with the mean values of traits GY, PHT, EHT and AD. Traits GY and PHT were constrained by $\mathbf{d}' = [0.5 \ -1.0]$ values. The Mean1 of the \hat{E}'_q value associated with traits GY and PHT (0.67 and -1.35, respectively) overestimated the \mathbf{d}' values by 35% (Table 2). However, note that the genetic gain per year, obtained as Mean2 = Mean1/1.5, underestimated the \mathbf{d}' values by 10% because the Mean2 of the \hat{E}'_q values associated with traits GY and PHT (0.45 and -0.90, respectively) were lower than the \mathbf{d}' values.

When we constrained three traits (GY, PHT and EHT) by vector $\mathbf{d}' = [0.5 \ -1.0 \ -0.5]$, we found results similar to those for two constraints. That is, the Mean1 of the \hat{E}'_q values associated with traits GY, PHT and EHT (0.75, -1.50 and -0.75, respectively) overestimated the \mathbf{d}' values by 50% (Table 2). However, for Mean2 = Mean1/1.5, we found that the \hat{E}'_q values associated with traits GY, PHT and EHT (0.50, -1.0 and -0.50, respectively) were equal to the \mathbf{d}' values. Thus, for dataset 1, the values of the vector of restrictions, $\mathbf{d}' = [0.5 \ -1.0]$ and $\mathbf{d}' = [0.5 \ -1.0 \ -0.5]$, were closer to Mean2 than to Mean1 values. This means that the estimated maximized expected genetic gain per trait estimated the genetic gain per year better than the genetic gain per selection cycle for dataset 1.

For dataset 2 ($p = 0.10$ and $k = 1.755$), with two ($\mathbf{d}' = [0.5 \ -1.0]$) and three ($\mathbf{d}' = [0.5 \ -1.0 \ -0.5]$) constraints, Mean1 for the \hat{E}'_q values associated with traits GY and PHT and those associated with GY, PHT and EHT, were closer to the \mathbf{d}' values than the Mean2 values (Table 2). In this case, the estimated expected genetic gain per trait estimated the genetic gain per year better than the genetic gain per selection cycle.

Dataset 1 gave higher estimated maximized CLGSI expected genetic gains per trait and genetic gains per year than dataset 2. This means that the number of genotypes affected the estimated values of the expected genetic gains (Table 1). The number of genotypes in real dataset 1 for cycles 0, 1 and 2 were 247, 320 and 303, respectively, whereas for dataset 2, the number of genotypes for those cycles were 181, 274 and 274 (Table 1), respectively.

Estimated maximized CLGSI selection response: The results of the estimated maximized CLGSI selection responses for both datasets and constraints, were as follows. For dataset 1, the Mean1 of the estimated maximized CLGSI selection responses for two and three constraints were 4.24 and 4.29, respectively, while for dataset 2, those estimated values were 2.79 and 3.01. For dataset 1, the Mean2 of the estimated maximized CLGSI selection responses for two and three constraints were 2.82 and 2.99, respectively, while for dataset 2, those estimated values were 1.86 and 2.01. These results indicate that dataset 1 gave higher estimated CLGSI genetic gains per year. Again, we explain these results by noting that the number of genotypes in real dataset 1 for cycles 0, 1 and 2 were 247, 320 and 303, respectively,

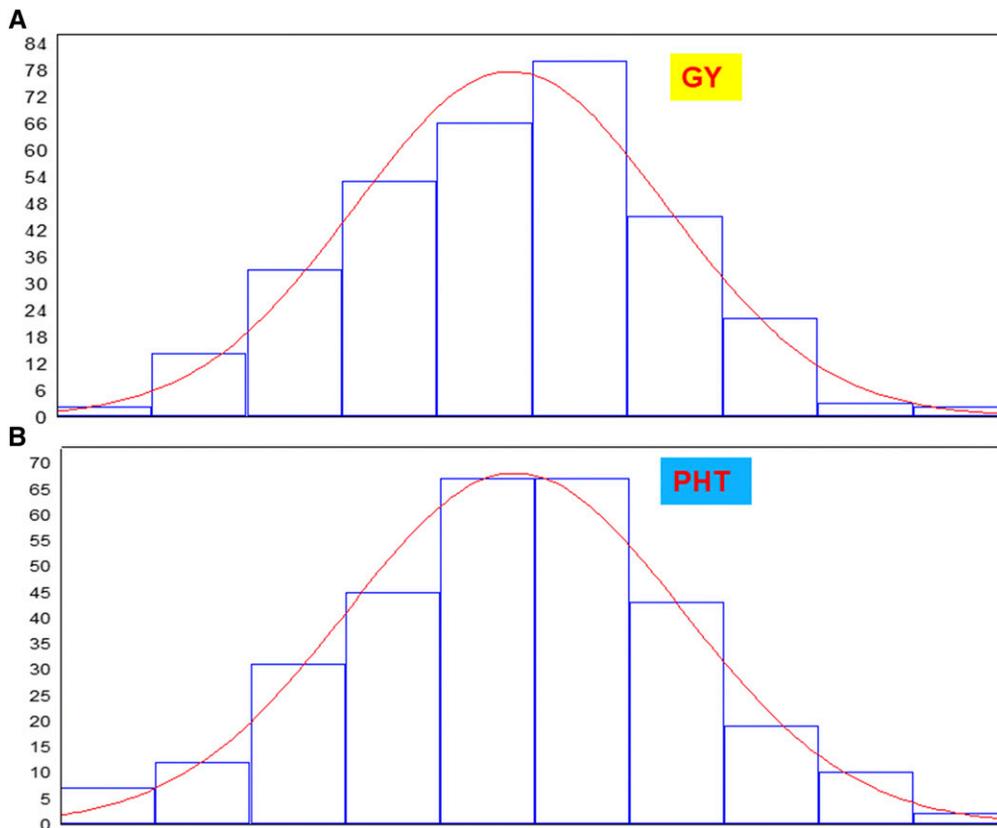


Figure 1 Distribution of the GEBVs (genomic estimated breeding values) associated with traits GY (Grain Yield, Fig. 1a) and PHT (Plant Height, Fig. 1b) for real Dataset 1 in cycle 1.

while for dataset 2, the number of genotypes for those cycles were 181, 274 and 274 (Table 1), respectively. This means that, in effect, the number of genotypes affected the estimated values of the selection response.

When we compared the CLGSI results (Table 2) with unconstrained LGSI results (Table 3), we found that the estimated CLGSI expected genetic gains per trait were different to the estimated unconstrained LGSI expected genetic gains per trait for both datasets. However, the estimated selection responses of both indices were very similar (Tables 2 and 3); this means that the two sets of constraints imposed on the CLGSI when we obtained its vector of coefficients mainly affected the CLGSI expected genetic gain per trait, as we would expect.

Simulated data

Correlation of the GEBV With the true breeding values of the traits: Figure 3 presents the estimated correlations between the GEBVs and true breeding values for four traits in six (C2 to C7) selection cycles. Each selection cycle contains four columns: the first column (from left to right) corresponds to the correlation between the GEBV and the T1 true breeding values; the second column corresponds to the correlation between the GEBV and the T2 true breeding values, and so on. In this figure, all correlation values tend to decrease. In C7, the correlation values between the GEBVs and the traits' true breeding values were 0.40, 0.55, 0.54, and 0.50 for each of the four traits, respectively, whereas in cycle two (C2), these correlations were 0.52, 0.74, 0.69, and 0.73 for each of the four traits, respectively. In percentage terms, the correlation values of C7 were only 76%, 74%, 78%, and 68% of the correlation values in C2. That is, the correlation between the GEBVs and the traits' true breeding values decreased more for traits 2 and 4 than for traits 1 and 3. We can explain these results by the number of QTL that

affected each trait and the size of the QTL effects on the traits in each selection cycle. In all selection cycles, the estimated correlations were higher than or equal to 0.45; thus, the GEBVs obtained with the simulated data were good predictors of the individual breeding values, and so the CLGSI was a good predictor of the net genetic merit because the CLGSI is a linear combination of GEBV.

Estimated maximized CLPSI and CLGSI parameters: Tables 4 and 5 present the estimated maximized CLPSI and CLGSI parameters, whereas Table 6 presents the unconstrained estimated maximized linear phenotypic and genomic selection index (LPSI and LGSI, respectively) parameters for seven simulated selection cycles. Céron-Rojas and Crossa (2018, Chapters 2 and 5) have given details of how to estimate the LPSI and LGSI parameters. We imposed two sets of constraints on the CLPSI and CLGSI expected genetic gains for seven selection cycles (C1 to C7). The vector for two constraints imposed on traits T1 and T2 was $\mathbf{d}' = [5.0 \quad -2.0]$, whereas the vector for three constraints imposed on traits T1, T2 and T3 was $\mathbf{d}' = [5.0 \quad -2.0 \quad 3.0]$. The estimated theta values were all positive for the simulated dataset in the seven selection cycles for both sets of constraints, which indicates that CLPSI and CLGSI move the population means in the desired direction.

We shall compare the estimated maximized CLPSI parameters to the estimated maximized CLGSI parameters using the results in Tables 4 and 5. In Tables 4 and 5, there are three means: Mean1 and Mean2 (Mean3) = Mean1/L (L = interval between selection cycles) associated with the estimated maximized CLPSI and CLGSI expected genetic gain per trait and selection response, respectively. Mean1 is the average of the seven selection cycles, whereas, for the CLPSI, Mean2 = Mean1/4 (4 = interval between selection cycles, L (Beyene *et al.* 2015)) is the

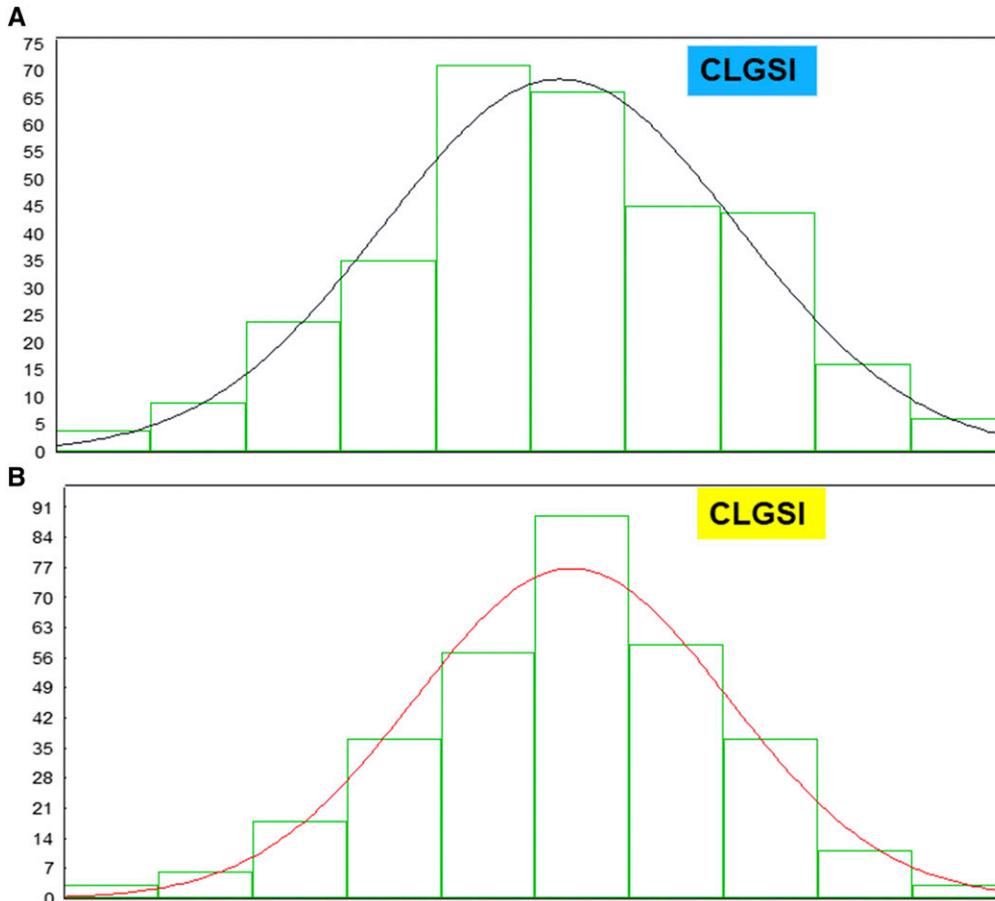


Figure 2 Distribution of the estimated CLGSI (constrained linear genomic selection index) values with two (Fig. 2a) and three (Fig. 2b) constraints for real Datasets 1 and 2, in cycles 1 and 2, respectively.

average genetic gain per year. For the CLGSI, $L = 1.5$ (Beyene *et al.* 2015), from where $\text{Mean3} = \text{Mean1}/1.5$. This means that the interval between selection cycles is a characteristic of each index and that the interval was shorter for the CLGSI than for the CLPSI.

The total proportion retained for both indices was $p = 0.10$ ($k = 1.755$) for all seven selection cycles. Each estimated CLPSI and CLGSI expected genetic gain per trait \hat{E}_q' ($q = 1, 2, \dots, 7$) for two and three constraints is a vector of values associated with the expected genetic values of traits T1, T2, T3 and T4 (Equation 3). Some authors (Ceron-Rojas *et al.* 2016; Céron-Rojas and Crossa 2018, Chapter 3) have described how to obtain \hat{E}_q' for the CLPSI. In **Appendix 2** (Equations A9 and A10), we describe how to estimate the maximized CLGSI selection response and expected genetic gain per trait.

Estimated and maximized CLPSI and CLGSI expected genetic gain per trait:

For two constraints ($\mathbf{d}' = [5 \ -2]$), the Mean1 of the estimated CLPSI expected genetic gain values associated with traits T1 and T2 were 6.99 and -2.80, respectively, whereas the Mean1 of the estimated CLGSI expected genetic gain values associated with traits T1 and T2 were 5.91 and -2.36, respectively. This means that the CLPSI overestimated the \mathbf{d}' values by 38.80%, whereas the CLGSI overestimated those values by 18.20%.

For three constraints ($\mathbf{d}' = [5.0 \ -2.0 \ 3.0]$), the Mean1 of the estimated CLPSI expected genetic gain values associated with traits T1, T2 and T3 were 5.86, -2.34 and 3.52, respectively, whereas the Mean1 of the estimated CLGSI expected genetic gains associated with those traits were 4.69, -1.87 and 2.81, respectively. This means that while the CLPSI

overestimated the \mathbf{d}' values by 17.20%, the CLGSI underestimated those values by only 6.20%. Thus, the expected genetic gain per trait per selection cycle of the CLGSI was closer to the true gain than was the expected gain of the CLPSI. In addition, because the interval between selection cycles was higher for the CLPSI ($L = 4$) than for the CLGSI ($L = 1.5$), the genetic gain per year associated with the expected genetic gain per trait was also higher for the CLGSI than for the CLPSI (Tables 4 and 5).

Finally, note that the estimated LPSI and LGSI expected genetic gains per trait (Table 6) were more different from the vectors of constraints, than the estimated CLPSI and CLGSI expected genetic gains per trait. Thus, the two sets of constraints imposed on the CLGSI (CLPSI) when we obtained its vector of coefficients, affected the CLGSI (CLPSI) expected genetic gain per trait, as we would expect.

Efficiency of the CLPSI and CLGSI selection response: In this subsection, we shall compare the average (Mean1) of the estimated CLPSI and CLGSI selection response to the true selection response obtained for both indices (Tables 4 and 5). In **Appendix 2** (Equations A11a and A11b), we indicate how we obtained the maximum true selection response for LPSI (LGSI) and for CLPSI (CLGSI). The maximum possible true selection responses for the economic index was the same for the genomic and phenotypic indices when we started the selection process, as this did not depend on the weights used in a particular selection index. However, from selection cycle two to selection cycle seven, the maximum possible true selection responses for both indices were different (Ceron-Rojas *et al.* 2015).

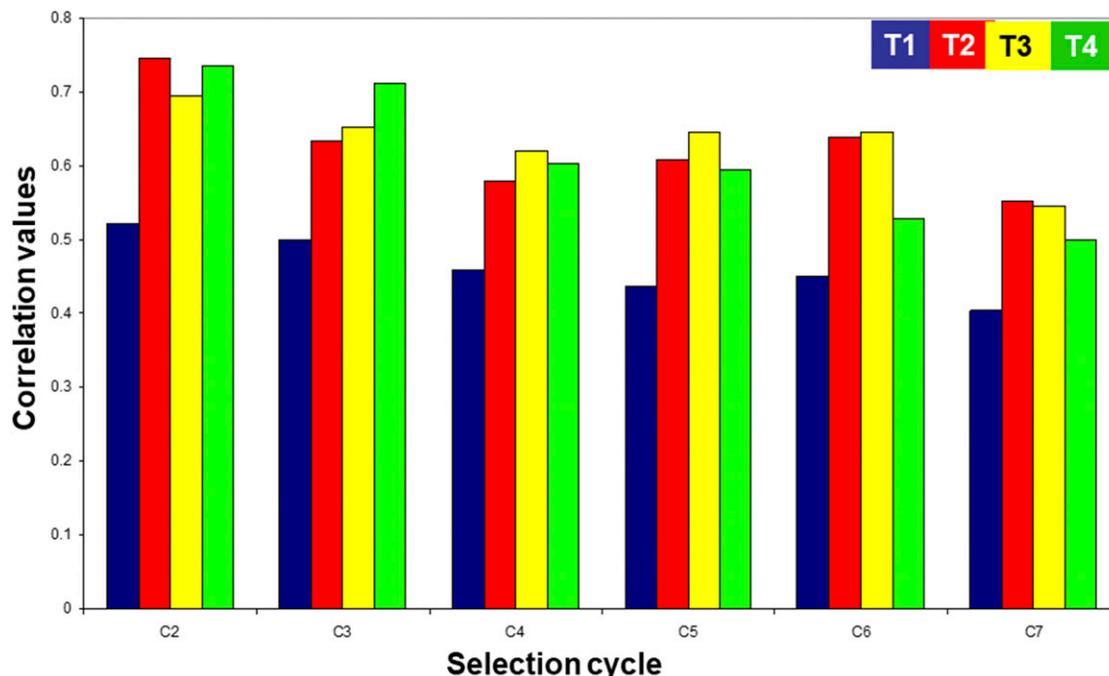


Figure 3 Correlations between the genomic estimated breeding values (GEBVs) and the true breeding values for four traits (T1, T2, T3 and T4) in six (C2 to C7) selection cycles. For each selection cycle, each column correspond to the correlations between the GEBV and the true breeding values for traits T1, T2, T3, and T4, respectively.

Again, let $p = 0.10$ ($k = 1.755$) be the total proportion retained for both indices for all seven selection cycles with two set of constraints, as indicated in the last subsection. For two constraints, the average of the estimated CLPSI selection response (13.91) explained 84.66% of the true selection response (16.43), while the average of the estimated CLGSI selection response (11.94) explained only 80.0% of the true selection response (14.89) (Table 4). In a similar manner, for three constraints, the average of the estimated CLPSI selection response (13.14) explained 89.02% of the true selection response (14.76), while the average of the estimated CLGSI selection response (10.63) explained only 79.0% of the true selection response (13.41) (Table 5). That is, in this case, the estimated CLPSI selection response was closer to its true response than the estimated CLGSI selection response was to its true response. Note, however, that because the interval between selection cycles was higher for the CLPSI ($L = 4$) than for the CLGSI ($L = 1.5$), the genetic gain per year associated with the selection response was higher for the CLGSI than for the CLPSI (Tables 4 and 5) for both constraints.

Finally, when we compared the estimated CLPSI and LPSI selection responses, we observed that they were very similar. The same was true when we compared the estimated CLGSI and LGSI selection responses (Tables 4 and 6). This means that the two sets of constraints imposed on the expected genetic gains of both indices did not affect the selection response of the indices, at least for this dataset.

DISCUSSION

CLGSI (CLPSI) expected genetic gain per trait and selection response

For the two real and simulated datasets, the CLGSI and CLPSI results indicated that CLGSI was more efficient than CLPSI per unit of time, but not per selection cycle, when the criterion for comparing the efficiency of both indices was the estimated selection response.

However, when the criterion used to compare the efficiency of both indices was the estimated expected genetic gain per trait, the CLGSI was more efficient than the CLPSI per unit of time and per selection cycle. In addition, we found that when we compared the estimated CLPSI selection response with the LPSI selection response, they were very similar. The same was true for CLGSI and LGSI when we compared their selection responses. This means that the two sets of constraints imposed on the CLGSI (CLPSI) when we obtained its vector of coefficients, mainly affected the estimated and maximized CLGSI (CLPSI) expected genetic gain per trait, not its estimated selection response.

Interval between selection cycles

For the two real and simulated datasets, the time required to conduct a selection cycle is a function of the time required to collect the data needed to estimate the index parameters. For CLPSI, the interval between selection cycles was 4, while for the CLGSI it was 1.5 (Beyene *et al.* 2015). The interval between selection cycles was the main factor that made CLGSI a more efficient index than CLPSI when we measured the efficiency of both indices by the genetic gains per year. It could also be the main justification for implementing genomic selection programs instead of phenotypic selection programs.

Accuracy of the GEBV and the CLGSI values

For the simulated datasets, the average of the accuracy of the GEBV and the estimated correlation of the CLGSI with the net genetic merit were higher than or equal to 0.5. The accuracy of the GEBV in cycle 2 was 0.67, while in cycle 7 it was 0.5. The average of the estimated correlation of the CLGSI with the net genetic merit (data not presented) for the seven selection cycles was 0.8 for two constraints and 0.72 for three constraints. Thus, the correlation between the estimated CLGSI and

the net genetic merit decreases when the number of constraints increased. Céron-Rojas and Crossa (2018; Chapters 5 and 6) have given a method for estimating the correlation of the CLGSI with the net genetic merit. In addition, Cerón-Rojas *et al.* (2016) found similar results in the CLPSI context. These results indicated the reliability of the GEBV for predicting the breeding value of the traits and of the CLGSI for predicting the net genetic merit; thus breeders can use the CLGSI as a good option for performing GS when there are genotyped individuals.

The multivariate normality assumption

Based on the normality assumption of the estimated CLGSI (CLPSI) and GEBV values, we developed and applied the CLGSI (CLPSI) to real and simulated data. The histograms of the GEBV and the CLGSI values indicated that these values approached the normal distribution. The multivariate normality distribution is very important for breeding plant and animal quantitative traits because these traits show continuous variability and are the result of many gene effects interacting among themselves and with the environment. That is, quantitative traits are the result of unobservable gene effects distributed across plant or animal genomes, which interact among themselves and with the environment to produce the observable characteristic plant and animal phenotypes (Falconer and Mackay 1996). Under the multivariate normal distribution assumption, the traits under selection can be described using only means, variances, and covariances. In addition, if the traits are not correlated, they are independent; linear combinations of traits are also normal; and even when the trait phenotypic values do not have that distribution, the normal distribution serves as a useful approximation, especially in inferences involving sample mean vectors, which, by the central limit theorem, have multivariate normal distribution (Rencher 2002). By this reasoning, the fundamental assumptions in this work were that the GEBVs have multivariate normal distribution, while the net genetic merit and the index have bivariate normal distribution. Under the latter assumption, the regression of the net genetic merit on any linear function of the phenotypic values was linear.

Criteria for evaluating the relative efficiency of the indices

We used four main criteria to compare CLGSI vs. CLPSI efficiency when performing genomic and phenotypic selection. Those criteria were the estimates of the theta values, the expected genetic gain per trait values, the selection response, and the interval between selection cycles. For real and simulated data, the estimated theta values were always positive, as we would expect. The estimated expected genetic gain per trait indicated how close these estimates were to the constraints imposed by the breeder for each trait, whereas the estimated selection response predicts the mean value of the net genetic merit in the progeny population. The interval between selection cycles is the time required to collect information to evaluate the index and complete one selection cycle. The four criteria were useful for evaluating and comparing the efficiency of both indices.

CONCLUSION

We compared the relative efficiency of a CLGSI vs. a CLPSI. We determined the efficiency of both indices based on four criteria using real and simulated datasets. In both type of datasets, we found that the CLGSI genetic gain per year was higher than the CLPSI genetic gain per year because the CLGSI interval between selection cycles was shorter than the CLPSI interval. Therefore, breeders should use the CLGSI when performing selection.

ACKNOWLEDGMENTS

We are thankful for the financial support provided by the Bill & Melinda Gates Foundation (for maize and wheat breeding programs), the CIMMYT CGIAR CRP (maize and wheat), as well the USAID projects. We acknowledge the financial support provided by the Foundation for Research Levy on Agricultural Products (FFL) and the Agricultural Agreement Research Fund (JA) in Norway through Research NFR Grant 267806.

LITERATURE CITED

- Börner, V., and N. Reinsch, 2012 Optimising multistage dairy cattle breeding schemes including genomic selection using decorrelated or optimum selection indices. *Genet. Sel. Evol.* 44: 1–11. <https://doi.org/10.1186/1297-9686-44-1>
- Beyene, Y., K. Semagn, S. Mugo, A. Tarekegne, R. Babu *et al.*, 2015 Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Sci.* 55: 154–163. <https://doi.org/10.2135/cropsci2014.07.0460>
- Ceron-Rojas, J. J., J. Crossa, V. N. Arief, K. Basford, J. Rutkoski, D. Jarquín, G. Alvarado, Y. Beyene, K. Semagn, and I. DeLacy, 2015 A genomic selection index applied to simulated and real data. *G3 (Bethesda)* 5: 2155–2164. <https://doi.org/10.1534/g3.115.019869>
- Céron-Rojas, J. J., and J. Crossa, 2018 *Linear selection indices in modern plant breeding*, Springer, Cham, the Netherlands., Available at <https://link.springer.com/book/10.1007/978-3-319-91223-3>, <https://doi.org/10.1007/978-3-319-91223-3>
- Céron-Rojas, J. J., J. Crossa, F. H. Toledo, and J. Sahagún-Castellanos, 2016 A predetermined proportional gains eigen selection index method. *Crop Sci.* 56: 2436–2447. <https://doi.org/10.2135/cropsci2015.11.0718>
- Dekkers, J. C. M., 2007 Prediction of response to marker-assisted and genomic selection using selection index theory. *J. Anim. Breed. Genet.* 124: 331–341. <https://doi.org/10.1111/j.1439-0388.2007.00701.x>
- Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*, Longman, New York.
- Harville, D. A., 1975 Index selection with proportionality constraints. *Biometrics* 31: 223–225. <https://doi.org/10.2307/2529722>
- Hazel, L. N., 1943 The genetic basis for constructing selection indexes. *Genetics* 8: 476–490.
- Itoh, Y., and Y. Yamada, 1987 Comparisons of selection indices achieving predetermined proportional gains. *Genet. Sel. Evol.* 19: 69–82. <https://doi.org/10.1186/1297-9686-19-1-69>
- Isik, F., J. Holland, and C. Maltecca, 2017 *Genetic Data Analysis for Plant and Animal Breeding*, Springer, Cham, the Netherlands. <https://doi.org/10.1007/978-3-319-55177-7>
- Kempthorne, O., and A. W. Nordskog, 1959 Restricted selection indices. *Biometrics* 15: 10–19. <https://doi.org/10.2307/2527598>
- Lorenz, A. J., S. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi *et al.*, 2011 Genomic selection in plant breeding: knowledge and prospects. *Adv. Agron.* 110: 77–123. <https://doi.org/10.1016/B978-0-12-385531-2.00002-5>
- Mallard, J., 1972 The theory and computation of selection indices with constraints: A critical synthesis. *Biometrics* 28: 713–735. <https://doi.org/10.2307/2528758>
- Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819–1829.
- Podlich, D. W., and M. Cooper, 1998 QU-GENE: a simulation platform for quantitative analysis of genetic models. *Bioinformatics* 14: 632–653. <https://doi.org/10.1093/bioinformatics/14.7.632>
- Rencher, A. C., 2002 *Methods of multivariate analysis*, Wiley, New York. <https://doi.org/10.1002/0471271357>
- Schott, J. R., 2005 *Matrix analysis for statistics*, Ed. 2nd. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Schaeffer, L. R., 2006 Strategy for applying genome-wide selection in dairy cattle. *J. Anim. Breed. Genet.* 123: 218–223. <https://doi.org/10.1111/j.1439-0388.2006.00595.x>

Tallis, G. M., 1985 Constrained selection. *Jpn. J. Genet.* 60: 151–155.

<https://doi.org/10.1266/jjg.60.151>

Togashi, K., C. Y. Lin, and T. Yamazaki, 2011 The efficiency of genome-wide selection for genetic improvement of net merit. *J. Anim. Sci.* 89: 2972–2980. <https://doi.org/10.2527/jas.2009-2606>

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91: 4414–4423. <https://doi.org/10.3168/jds.2007-0980>

Communicating editor: J. Holland

APPENDIX 1

Genomic breeding values

In the training population, the i^{th} phenotypic value (y_i) can be denoted as $y_i = g_i + e_i$, where g_i is the breeding value (*i.e.*, the average additive effects of the genes an individual receives from both parents) and e_i is the residual. It is assumed that g_i and e_i are independent and that both have normal distribution with an expectation equal to zero and variance $\sigma_{g_i}^2$ and $\sigma_{e_i}^2$, respectively. The vector of genomic breeding values $\mathbf{z}'_i = [z_{i1} \ z_{i2} \ \dots \ z_{in}]$ associated with the vector of observations $\mathbf{y}'_i = [y_{i1} \ y_{i2} \ \dots \ y_{in}]$ (n = number of observations for trait i , $i = 1, 2, \dots, t$; t = number of traits) of the candidates for selection for trait i can be written as

$$\mathbf{z}_i = \mathbf{X}\mathbf{u}_i, \quad (\text{A1})$$

where \mathbf{X} is an $n \times m$ matrix (n = number of observations and m = number of markers in the population) of coded marker values ($2 - 2q$, $1 - 2q$ and $-2q$ for genotypes AA , Aa , and aa , respectively) associated with the additive effects of the quantitative trait loci (QTL), and \mathbf{u}_i is an $m \times 1$ vector of the additive effects of the QTL associated with markers that affect the i^{th} trait. It is assumed that \mathbf{z}_i has multivariate normal distribution (MVN) with mean $\mathbf{0}$ and covariance matrix $\mathbf{G}\sigma_{z_i}^2$, where $\sigma_{z_i}^2$ is the additive genomic variance of \mathbf{z}_i and $\mathbf{G} = \mathbf{X}\mathbf{X}'/c$ is the $n \times n$ additive genomic relationship matrix; $c = \sum_{j=1}^m 2q_j(1 - q_j)$ in an F_2 population; q_j is the frequency of allele A and $1 - q_j$ is the frequency of allele a in the j^{th} marker ($j = 1, 2, \dots, m$).

Let g_{ij} ($i = 1, 2, \dots, t$, t = number of traits; $j = 1, 2, \dots, n$, n = number of observations) be the j^{th} element of the vector of true genotypic breeding values $\mathbf{g}'_i = [g_{i1} \ g_{i2} \ \dots \ g_{in}]$; then the covariance between z_{ij} and g_{ij} is $\text{Cov}(g_{ij}, z_{ij}) = \sigma_{z_i}^2$. That is, the covariance between z_{ij} and g_{ij} is equal to the additive genomic variance of z_{ij} (Dekkers 2007). Let $\mathbf{z}' = [z_1 \ z_2 \ \dots \ z_t]$ and $\mathbf{g}' = [g_1 \ g_2 \ \dots \ g_t]$ be $1 \times t$ vectors of genomic and true breeding values, respectively, for t traits; then

$$\mathbf{\Gamma} = \text{Var}(\mathbf{z}) = \text{Cov}(\mathbf{z}, \mathbf{g}) \quad (\text{A2})$$

is the covariance matrix of genomic breeding values.

Estimating marker effects

Let $\mathbf{u}' = [\mathbf{u}'_1 \ \mathbf{u}'_2 \ \dots \ \mathbf{u}'_t]$ be a vector $1 \times nt$ associated with t traits. The i^{th} vector \mathbf{u}_i ($i = 1, 2, \dots, t$) of marker effects in the training population can be estimated as $\hat{\mathbf{u}}_i = c^{-1}\mathbf{X}'[\mathbf{G} + \nu\mathbf{I}_n]^{-1}(\mathbf{y}_i - \mathbf{1}\mu_i)$, where $\nu = \frac{\sigma_{e_i}^2}{\sigma_{g_i}^2}$, $\sigma_{g_i}^2$, $\sigma_{e_i}^2$, and the other parameters were defined earlier. In addition, we can estimate vector $\mathbf{u}' = [\mathbf{u}'_1 \ \mathbf{u}'_2 \ \dots \ \mathbf{u}'_t]$ in the multi-trait context as

$$\hat{\mathbf{u}} = c^{-1}\mathbf{W}'[(\mathbf{I}_t \otimes \mathbf{G}) + (\mathbf{N} \otimes \mathbf{I}_n)]^{-1}(\mathbf{y} - \boldsymbol{\mu} \otimes \mathbf{1}), \quad (\text{A3})$$

where $\mathbf{W} = \mathbf{I}_t \otimes \mathbf{X}$, “ \otimes ” denotes the Kronecker product (Schott 2005); c and \mathbf{X} were defined in Equation (A1); $\mathbf{N} = \mathbf{R}\mathbf{C}^{-1}$, \mathbf{R} and \mathbf{C} are the residual and genotypic covariance matrices for t traits, respectively; $\mathbf{y}' = [y'_1 \ y'_2 \ \dots \ y'_t] \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{V})$ is a vector of size $1 \times tn$, with covariance matrix $\mathbf{V} = \mathbf{C} \otimes \mathbf{G} + \mathbf{R} \otimes \mathbf{I}_n$; \mathbf{I}_t and \mathbf{I}_n are identity matrix of size $t \times t$ and $n \times n$, respectively; $\boldsymbol{\mu}' = [\mu_1 \ \mu_2 \ \dots \ \mu_t]$ is a vector $1 \times t$ of means associated with vector \mathbf{y} , and $\mathbf{1}$ is a vector $n \times 1$ of 1's.

Estimating genomic breeding values

We can estimate the vector $\mathbf{z}' = [z'_1 \ z'_2 \ \dots \ z'_t]$ $1 \times nt$ of genomic breeding values for t traits in the testing population as

$$\hat{\mathbf{z}} = \mathbf{W}\hat{\mathbf{u}}. \quad (\text{A4})$$

Equation (A4) is the vector of GEBV for the multi-trait case and $\mathbf{W} = \mathbf{I}_t \otimes \mathbf{X}$ (Equation A3). Thus, in the testing population, the only thing that will change in Equation (A4) will be the coded values in matrix \mathbf{X} , while $\hat{\mathbf{u}}$ will be the same in each selection cycle. In Equations (A3) and (A4), we assumed that $\boldsymbol{\mu}$, \mathbf{C} and \mathbf{R} are known.

Estimating the genomic covariance matrix $\mathbf{\Gamma}$

Let $\hat{\mathbf{z}}_j = \mathbf{X}\hat{\mathbf{u}}_j$ and $\hat{\mathbf{z}}_i = \mathbf{X}\hat{\mathbf{u}}_i$ (Equation A4) be the genomic estimated breeding values (GEBVs) of $\mathbf{z}_j = \mathbf{X}\mathbf{u}_j$ and $\mathbf{z}_i = \mathbf{X}\mathbf{u}_i$ (Equation A1), respectively, and denote by $\hat{\mu}_j$ and $\hat{\mu}_i$ the arithmetic means of the values of $\hat{\mathbf{z}}_j$ and $\hat{\mathbf{z}}_i$. We can estimate matrix $\mathbf{\Gamma}$ (Equation A2) when there is no phenotypic information, as

$$\hat{\mathbf{\Gamma}} = \{\hat{\sigma}_{ji}\}, \quad (\text{A5})$$

where $\hat{\sigma}_{ji} = \frac{1}{g}(\hat{\mathbf{z}}_j - \mathbf{1}\hat{\mu}_j)' \mathbf{G}^{-1}(\hat{\mathbf{z}}_i - \mathbf{1}\hat{\mu}_i)'$ is the covariance between $\hat{\mathbf{z}}_j = \mathbf{X}_i\hat{\mathbf{u}}_j$ and $\hat{\mathbf{z}}_i = \mathbf{X}\hat{\mathbf{u}}_i$ values ($j, i = 1, 2, \dots, t$); g is the number of genotypes; $\mathbf{1}$ is a $g \times 1$ vector of 1's and $\mathbf{G} = c^{-1}\mathbf{X}\mathbf{X}'$ is the additive genomic relationship matrix (Ceron-Rojas *et al.*, 2015, for details).

APPENDIX 2

The CLGSI vector of coefficients

Let $\mathbf{D}' = \begin{bmatrix} d_r & 0 & \dots & 0 & -d_1 \\ 0 & d_r & \dots & 0 & -d_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & d_r & -d_{r-1} \end{bmatrix}$ be a Mallard (1972) matrix $(r - 1) \times r$ of constrained values, where d_i is the i^{th} element of vector

$\mathbf{d}' = [d_1 \ d_2 \ \dots \ d_r]$ (r is the number of traits constrained) of constrained values imposed by the breeder on the expected genetic gain of the trait (Equation 3), and let \mathbf{U}' be a Kempthorne and Nordskog (1959) matrix $(t - r) \times t$ (t = number of traits) of 1's and 0's described in **Appendix 3**. To obtain the CLGSI vector of coefficients ($\boldsymbol{\beta}$) that maximizes Equations (2) and (3), we will minimize the mean squared difference between $H = \mathbf{w}'\mathbf{g}$ and $I = \boldsymbol{\beta}'\mathbf{z}$ ($E[(H - I)^2]$) with respect to $\boldsymbol{\beta}$ under the restriction $\mathbf{D}'\mathbf{U}'\boldsymbol{\Gamma}\boldsymbol{\beta} = \mathbf{0}$.

Let $\mathbf{M}' = \mathbf{D}'\mathbf{U}'\mathbf{\Gamma}$ and suppose that $\mathbf{\Gamma}$, \mathbf{U}' , \mathbf{D}' and \mathbf{w} are known. To minimize $E[(H-I)^2]$ under the restriction $\mathbf{M}'\mathbf{\beta} = \mathbf{0}$, we need to minimize the function

$$\Phi(\mathbf{\beta}, \mathbf{v}) = \mathbf{\beta}'\mathbf{\Gamma}\mathbf{\beta} - 2\mathbf{w}'\mathbf{\Gamma}\mathbf{\beta} + 2\mathbf{v}'\mathbf{M}'\mathbf{\beta} \quad (\text{A6})$$

with respect to vectors $\mathbf{\beta}$ and $\mathbf{v}' = [v_1 \ v_2 \ \dots \ v_{r-1}]$, where \mathbf{v} is a vector of Lagrange multipliers. Equation (A6) derivative results are:

$$\begin{bmatrix} \mathbf{\Gamma} & \mathbf{M}' \\ \mathbf{M}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{\beta} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{\Gamma}\mathbf{w} \\ \mathbf{0} \end{bmatrix}, \text{ from where the vector that minimizes } E[(H-I)^2] \text{ under the restriction } \mathbf{M}'\mathbf{\beta} = \mathbf{0} \text{ is}$$

$$\mathbf{\beta} = \mathbf{K}\mathbf{w}, \quad (\text{A7})$$

where $\mathbf{K} = [\mathbf{I}_t - \mathbf{Q}]$, $\mathbf{Q} = \mathbf{U}\mathbf{D}(\mathbf{D}'\mathbf{U}'\mathbf{\Gamma}\mathbf{U}\mathbf{D})^{-1}\mathbf{D}'\mathbf{U}'\mathbf{\Gamma}$ and \mathbf{I}_t is an identity matrix of size $t \times t$. If in Equation (A7), $d_i = 0$ for all traits, then $\mathbf{D} = \mathbf{U}$ and the CLGSI will be similar to the Kempthorne and Nordskog (1959) index, but in the genomic selection context. When $\mathbf{D} = \mathbf{U}$ and \mathbf{U}' is a null matrix, $\mathbf{\beta} = \mathbf{w}$, the vector of economic weights or the unconstrained LGSI vector of coefficients (Ceron-Rojas *et al.*, 2015). Thus, the CLGSI is more general than the LGSI and includes the LGSI as a particular case.

Estimating the CLGSI vector of coefficients

We can estimate $\mathbf{\beta} = \mathbf{K}\mathbf{w}$ (Equation A7) as

$$\hat{\mathbf{\beta}} = \hat{\mathbf{K}}\mathbf{w}, \quad (\text{A8})$$

where $\hat{\mathbf{K}} = [\mathbf{I}_t - \hat{\mathbf{Q}}]$, $\hat{\mathbf{Q}} = \mathbf{U}\mathbf{D}(\mathbf{D}'\mathbf{U}'\hat{\mathbf{\Gamma}}\mathbf{U}\mathbf{D})^{-1}\mathbf{D}'\mathbf{U}'\hat{\mathbf{\Gamma}}$. We defined all the parameters of Equation (A8) earlier.

Estimating the maximized CLGSI selection response and expected genetic gain per trait

The estimated maximized CLGSI selection response (Equation 5) is

$$\hat{R} = k\sqrt{\hat{\mathbf{\beta}}'\hat{\mathbf{\Gamma}}\hat{\mathbf{\beta}}}, \quad (\text{A9})$$

while the estimated optimized CLGSI expected genetic gain per trait (Equation 3) is

$$\hat{E} = k\frac{\hat{\mathbf{\Gamma}}\hat{\mathbf{\beta}}}{\sqrt{\hat{\mathbf{\beta}}'\hat{\mathbf{\Gamma}}\hat{\mathbf{\beta}}}}. \quad (\text{A10})$$

Maximized true selection responses

We obtained the maximized true selection response for the unconstrained linear phenotypic and genomic selection indices (LPSI and LGSI, respectively) as

$$R_p = k\sqrt{\mathbf{w}'\mathbf{C}\mathbf{w}} \quad \text{and} \quad R_G = k\sqrt{\mathbf{w}'\mathbf{\Gamma}\mathbf{w}}, \quad (\text{A11a})$$

respectively, where matrices \mathbf{C} and $\mathbf{\Gamma}$ are the covariance matrix of true genotype values associated with the phenotypic (\mathbf{C}) and the genomic values ($\mathbf{\Gamma}$, Equation A2) in each selection cycle, k is the selection intensity and \mathbf{w} is a vector of economic weights. In addition, the maximized true selection responses for CLPSI and CLGSI are

$$R_p = k\sqrt{\mathbf{b}_C'\mathbf{C}\mathbf{b}_C} \quad \text{and} \quad R_G = k\sqrt{\mathbf{b}_\Gamma'\mathbf{\Gamma}\mathbf{b}_\Gamma}, \quad (\text{A11b})$$

respectively, where $\mathbf{b}_C = \mathbf{K}_C\mathbf{w}$ and $\mathbf{b}_\Gamma = \mathbf{K}_\Gamma\mathbf{w}$ are the vector of coefficients; $\mathbf{K}_C = [\mathbf{I}_t - \mathbf{Q}_C]$, $\mathbf{K}_\Gamma = [\mathbf{I}_t - \mathbf{Q}_\Gamma]$, $\mathbf{Q}_C = \mathbf{U}\mathbf{D}(\mathbf{D}'\mathbf{U}'\mathbf{C}\mathbf{U}\mathbf{D})^{-1}\mathbf{D}'\mathbf{U}'\mathbf{C}$, and $\mathbf{Q}_\Gamma = \mathbf{U}\mathbf{D}(\mathbf{D}'\mathbf{U}'\mathbf{\Gamma}\mathbf{U}\mathbf{D})^{-1}\mathbf{D}'\mathbf{U}'\mathbf{\Gamma}$ (Equation A7). Matrices \mathbf{C} and $\mathbf{\Gamma}$ were defined in Equation (A11a). We obtained these last parameters only for the simulated datasets.

APPENDIX 3

The \mathbf{U}' matrix of restrictions

In addition to matrix \mathbf{D} , to obtain the CLGSI (CLPSI) vector of coefficients, we need the matrix \mathbf{U}' ($t - r$) \times t ($t =$ number of traits; $r =$ number of traits restricted), which is a matrix of 1's and 0's (1 indicates that the trait is restricted and 0 that the trait is not restricted), and depends on the number of restricted traits. Thus, suppose that we restrict only one of t traits; then, we can restrict the first of them as $\mathbf{U}' = [1 \ 0 \ 0 \ \dots \ 0]$, the second as $\mathbf{U}' = [0 \ 1 \ 0 \ \dots \ 0]$, the third as $\mathbf{U}' = [0 \ 0 \ 1 \ \dots \ 0]$, etc. When we restrict two of t traits, matrix \mathbf{U}' could be constructed as follows.

We can restrict the first and second traits as $\mathbf{U}' = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \end{bmatrix}$, the first and third traits as $\mathbf{U}' = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \end{bmatrix}$, the second and third traits as $\mathbf{U}' = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \end{bmatrix}$, etc. If we restrict three of t traits, matrix \mathbf{U}' will have the following form when the first, second and

third traits are restricted: $\mathbf{U}' = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \end{bmatrix}$; if the first, second and fourth traits are restricted, $\mathbf{U}' = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \end{bmatrix}$, if the

second, the third and the fourth traits are restricted, $\mathbf{U}' = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \end{bmatrix}$, etc. The procedure used to construct matrix \mathbf{U}' is valid for any

number of restricted traits (Ceron-Rojas and Crossa 2018, Chapter 3, for details).