CROP BREEDING AND APPLIED BIOTECHNOLOGY

ARTICLE

Population structure correction for genomic selection through eigenvector covariates

Camila Ferreira Azevedo[1*], Marcos Deon Vilela de Resende[1,2], Fabyano Fonseca e Silva[3], Moysés Nascimento[1], José Marcelo Soriano Viana[4] and Magno Sávio Ferreira Valente[4]

Crop Breeding and Applied Biotechnology
17: 350-358, 2017
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332017v17n4a53

Abstract: *We proposed a population structure correction for genome-wide selection based on covariance analysis via eigenvector (EVG) decomposition. The agreement between the predicted and true breeding values was evaluated by independent cross-validation data sets. Other correction methods such as correction via principal components, best linear unbiased prediction, and deregressed phenotype were also evaluated. Based on different simulation scenarios, the proposed EVG out performed the other methods in the prediction of accuracy.*

Key words*: Short-term genomic prediction, Mendelian segregation, principal component, deregressed phenotype.*

INTRODUCTION

In genome-wide selection (GWS), the correction for population structure allows eliminating the effects of genitors or families from the total estimated breeding values in order to obtain the effects of pure Mendelian segregation. Thus, association analyses between allele markers and quantitative trait loci (QTLs) capture the genetic effects consequences of linkage disequilibrium (LD) by excluding relationship-based genealogy.

The performance of short-term genomic prediction (at present or subsequent generations) depends on genealogical kinship more than on LD (Resende et al. 2012). Therefore, predictions based on total genetic values, which consider both kinds of information, are recommended. On the other hand, long-term genomic prediction (prediction for future generations) only requires LD information between markers and QTLs, which is called Mendelian segregation effect. Thus, correction for population structure is required to obtain accurate predictions (Resende et al. 2012). Structure correction is also important for genome-wide association studies (GWAS) because they aim to detect causative variants and allow for association analysis between alleles and QTLs by considering only LD (free genealogy) and avoiding the occurrence of false positives (Daetwyler et al. 2012).

Therefore, statistical methods for population structure correction are required for efficient long-term GWS. Among them, correction via best linear unbiased prediction (BLUP) (Resende et al. 2012) and procedures based on deregressed phenotype (DP) (Garrick et al. 2009) are highlighted. Covariance analysis via eigenvectors (EVG), which is based on the genomic relationship matrix and was

***Corresponding author:**
E-mail: camila.azevedo@ufv.br

Received: 02 May 2016
Accepted: 03 May 2017

[1] Universidade Federal de Viçosa (UFV), Departamento de Estatística, 36.570-900, Viçosa, MG, Brazil
[2] UFV, Departamento de Engenharia Florestal
[3] UFV, Departamento de Zootecnia
[4] UFV, Departamento de Biologia Geral

presented by Patterson et al. (2006) and Yang et al. (2011) to correct for population structure in GWAS, has never been tested for genomic prediction. The inclusion of principal components (PC) (Daetwyler et al. 2012, Riedelsheimer et al. 2012, Price et al. 2006) obtained from the markers used in the model has also been proposed for population structure correction in both GWAS and GWS. The abovementioned correction methods consist of including eigenvectors from the genomic relationship matrix G in the model as fixed covariates. The eigenvectors are associated with the highest eigenvalues and first PCs of G. Thus, since G is linked to individuals and not to markers, the objective is to capture genetic variations depending on population structure. However, from the viewpoint of genomic selection, no reports on the comparison among these four population structure correction methods have been published yet.

Based on previous evaluations of methodologies for GWS, GWAS, and QTL mapping conducted with simulation studies (Azevedo et al. 2015, Piccoli et al. 2014, Ribeiro et al. 2005), we assessed the performance of the EVG method or population structure correction in terms of efficiency in estimating the free genomic values of relationships using simulated data under different scenarios (i.e., 2 heritability levels × 2 genetic architectures). We compared the proposed EVG method with other methods (i.e., BLUP, DP, and PC) in terms of prediction accuracy obtained from each scenario.

MATERIAL AND METHODS

Simulated datasets

The datasets generated according to Azevedo et al. (2015) were simulated using the Real Breeding software (Viana 2013). Two random mating populations in linkage equilibrium were crossed to generate a population of 5,000 elements from 100 families using LD, which was subjected to five generations of random mating without mutation, selection, or migration. The resultant population is an advanced generation composite, which presents Hardy-Weinberg equilibrium and LD. According to Viana (2004), the LD value (Δ) in a composite population is $\Delta_{ab} = \left(\frac{1-2\theta_{ab}}{4}\right)(p_a^1 - p_a^2)(p_b^1 - p_b^2)$, where *a* and *b* are two single nucleotide polymorphisms (SNPs), two QTLs, or one SNP and one QTL, θ is the frequency of recombinant gametes, and p^1 and p^2 are the allele frequencies in the parental populations 1 and 2, respectively. Moreover, the LD value depends on the allele frequencies in the parental populations. Thus, regardless of the distance between the SNPs and/or QTLs, if the allele frequencies are equal in the parental population, Δ = 0. The LD is maximized (|Δ| = 0.25) when θ = 0 and $|p^1 - p^2| = 1$. In this case, the LD value is positive with coupling and negative with repulsion (Azevedo et al. 2015, Viana 2004).

From the advanced generation of the composite, 1,000 individuals with diploid genomes having a length of 200 centimorgans (cM) (L=2 Morgans) were generated assuming 10equally sized chromosomes, each one with two haplotypes. Azevedo et al. (2015) simulated a marker density by assigning 2,000 equidistant SNP markers that were separated by 0.1 cM across the 10chromosomes. Interestingly, 100 of the 2,000 markers were actually genes (QTL). The 1,000 individuals that came from the same generation and from 20 full-sib families (each one with 50 individuals) were genotyped and phenotyped. This simulation provided a typical, small effective population size (*Ne*= 39.22) and large LD in the breeding populations. *Ne* of approximately 40 and 50 individuals per family are typical values in elite breeding populations of allogamous plant species (Resende 2002).

The QTLs were distributed among the regions covered by the SNPs. For each trait, we informed the degree of dominance (*d*/*a*, where *a* and *d* are the genotypic values for one homozygote and heterozygote, respectively) and the direction of dominance (positive and/or negative). The obtained genotypic values for homozygotes were within the limits of *Gmax=100(m + a)* and *Gmin=100(m - a)* where *m* is the mean of genotypic values, which are the maximum and minimum values, respectively. Goddard et al. (2011) presented the realized proportion (r^2_{mq}) of genetic variation explained by the markers as $r^2_{mq} = \frac{n}{n+n_{QTL}}$, where n_{QTL} is the number of QTL and *n* is the number of markers. With *n* = 2,000 markers and n_{QTL}= 100, we obtained r^2_{mq} = 0.95, which revealed that the genome was sufficiently saturated by markers.

Traits with two genetic architectures were simulated, one following the infinitesimal model and the other containing five major effect genes accounting for 50% of the genetic variability. For the former, one additive effect of small magnitude on the phenotype was assigned (under the normal distribution setting) to each of the 100 QTLs. For the

latter, small additive effects were assigned to the remaining 95 loci. The effects were normally distributed with zero mean and genetic variance (size of genetic effects) allowing the desired heritability level. The phenotypic value was obtained by adding to the genotypic value a random deviate from a normal distribution N (0, σ_e^2), where the variance σ_e^2 was defined according to two levels of narrow-sense heritabilities of approximately 0.20 and 0.35. Heritability levels were chosen to represent one trait with low heritability and another with moderate heritability, which addressed the cases where genomic selection is expected to prevail over phenotypic selection (Azevedo et al. 2015). The magnitudes of the narrow-sense and broad-sense heritability are associated with an average degree of dominance level (*d/a*) of approximately 1 (complete dominance) in a population with intermediate allele frequencies. Simulations assumed independence of additive and dominance effects, with dominance effects having the same distribution as the additive effects (both were normally distributed with zero mean). In the simulation, it was also observed that marker alleles had minor allele frequency (MAF) greater than 5%.

Scenarios

Four different scenarios were used in the analyses: two broad-sense heritability levels (around 0.20 and 0.35) × two genetic architectures (polygenic and mixed inheritance). The scenarios were defined as follows: Scenario 1, trait controlled by genes with small effects and heritability of 0.22; Scenario 2, trait controlled by genes with small effects and heritability of 0.37; Scenario 3, trait controlled by small and major effects genes and heritability of 0.20; Scenario 4, trait controlled by small and major effects of genes and heritability of 0.32. Each kind of population (or scenario) was simulated 10 times (Calus et al. 2008).

Eigenvectors correction

The EVG correction proposed in this study considers the eigenvectors of the genomic relationship matrix *G* as fixed covariates in the genomic mixed model. The eigenvectors are associated with the highest eigenvalues and first PCs of G. Thus, since G is linked to the individuals, and not to the markers, the purpose was to capture the genetic variance due to population structure.

The traditional phenotypic BLUP has been habitually applied in plant breeding to predict the genetic value (Resende and Barbosa 2006, Carvalho et al. 2008, Oliveira et al. 2011, Silva et al. 2015). If the relationship matrix A is computed via markers information (*G* matrix) and used within the traditional phenotypic BLUP, the method is called genomic best linear unbiased predictor (GBLUP). Thus, the GBLUP method was used and the eigenvectors from the G matrix were included in the model as follows:

$$y = Xb + \sum_{i=1}^{v} U_i a_i + Zg + e$$

where y is the vector of phenotypes, *v* is the number of eigenvectors U_i (here, 1-50) associated to the PCs with the highest eigenvalues, b is a vector of the other fixed effects with an incidence matrix *X*; *g* is a vector of additive random genetic effects of individuals, $g \sim N(0, G\sigma_g^2)$ where *G* is the genomic additive relationship matrix $G = \frac{MM'}{tr\left(\frac{MM'}{N}\right)}$ (*N* is the number of individuals, *tr* the trace operator matrix) and σ_g^2 is the additive genetic variance; and *e* is the residuals vector, $e \sim N(0, I\sigma_e^2)$ where I is the identity matrix and σ_e^2 the residual variance. The eigenvectors were fitted as fixed effects, in which a_i are the regression coefficients associated with them. The marker effects can be obtained by $\hat{m} = (M'M)^{-1}M'\hat{g}$, *M* being the marker incidence matrix with values of 0, 1, and 2.

Deregressed phenotype

The traditional method used for population structure correction in genomic selection was proposed by Garrick et al. (2009) and it is called DP or adjusted phenotype. It assumes that the genealogical information, variance components, and genetic values are known and it requires a mixed model analysis based on the relationship matrix via pedigree [i.e., individual model (IM)]. This correction is presented as follows:

$$\begin{bmatrix} Z'_{gm}Z_{gm} + 4\lambda^* & -2\lambda^* \\ -2\lambda^* & Z'_iZ_i + 2\lambda^* \end{bmatrix} \begin{bmatrix} \hat{g}_{gm} \\ \hat{g}_i \end{bmatrix} = \begin{bmatrix} Y_{gm} \\ Y_i \end{bmatrix},$$

Where $\hat{g}_i$ is the genetic value of the genotyped individual (i=1,...,N), $\lambda^* = \frac{1-h^2}{h^2}$, $\hat{g}_{gm} = \frac{\hat{g}_h + \hat{g}_k}{2}$ is the average genetic value of the h and k genitors, $Z'_{gm}Z_{gm}$ is the information content associated with the genitors average, Z'_iZ_i is the information content associated with the individuals including their progeny, and y_{gm} and y_i are the phenotypic observations corrected for fixed and individual effects, respectively.

The information matrix associated with the genitor average is unknown, and can be obtained by means of the following expression:

$$Z'_{gm}Z_{gm} = \lambda^*(0.5\gamma - 4) + 0.5\lambda^* \left(\gamma^2 + \frac{16}{\delta}\right)^{1/2}$$

where $\gamma = \frac{1}{0.5 - r^2_{gm}}$ and $\delta = \frac{0.5 - r^2_{gm}}{1 - r^2_i}$ with $r^2_{gm} = \frac{r^2_{gh} + r^2_{gk}}{2}$ being the reliability of the genetic value predicted for genitors h and k given by and r^2_i is the reliability associated with the predicted genetic value of the individual. Therefore, the information matrix associated with the individual Z'_iZ_i can be obtained through $Z'_{gm}Z_{gm}$. Thus, Z'_iZ_i is given by $\delta Z'_{gm}Z_{gm} + 2\lambda^*(2\delta - 1)$.

The phenotype vectors corrected for fixed and genitor effects are the output of the equations $\hat{y}_i = (-2\lambda^*)\hat{y}_{gm} + (Z'_iZ_i + 2\lambda^*)\hat{g}_i$. Thus, the phenotype will also be corrected for the average genetic value of its genitors. After this correction, the deregressed genetic value ($\hat{g}_i$) is obtained by $\hat{g}_i = \frac{\hat{y}_i}{Z'_iZ_i}$. Analogous to EVG, the marker effects can be obtained using the GBLUP method as shown earlier in the text (i.e., $\hat{m} = (M'M)^{-1}M'\hat{g}^*$).

Correction via BLUP

Population structure correction can be done by means of the additive genetic model proposed by Resende et al. (2012). The following model is considered:

$$y = Xh + Xg + e$$

where y is the vector of phenotypic records, b is the vector of the fixed effects with an incidence matrix X, g is the vector of additive random genetic effects of individuals, and e is the residuals vector. The correction via the computational model BLUP assumes that the individual genealogical information is known, thus, after the genetic values prediction for each individual and their respective genitors, the corrected genetic value for the genitors can be obtained by:

$$\hat{g}_c = \hat{g} - 0.5\hat{g}_h - 0.5\hat{g}_k$$

where $\hat{g}_h$ and $\hat{g}_k$ are the predicted genetic $\hat{g}_c$ values of the h and k genitors, respectively. The genetic value deregressed and corrected for the genitor effects, i.e., the effect of the Mendelian segregation, is given by:

$$\hat{g}^* = \frac{\hat{g}_c}{h^2_{sm}},$$

where $h^2_{sm} = \frac{0.5h^2}{0.5h^2 + (1 - h^2)}$ is the heritability of the Mendelian segregation and h^2 the heritability of the trait. This step is necessary because the genetic values do not undergo two regressions, i.e., one based on the relationship matrix via pedigree (mixed model analysis) and another one on the marker matrix. In this context, $\hat{g}^*$ is equivalent to the residual ($y - X\hat{b} - Z_{h,k}\hat{g}_{h,k}$) deriving from the reduced individual model, in which $\hat{g}_{h,k}$ is the predicted genetic value of the genitors.

After the above mentioned population structure correction, the marker effects can be estimated by means of the GBLUP model:

$$y_c = Zg^{**} + e$$

where $y_c = \hat{g}^*, g^{**}$ is the vector of additive random genetic effects of individuals with incidence matrix Z, which is here pre-corrected for population structure, and $g^{**} \sim N(0, G\sigma^2_g)$ with G being the genomic additive relationship matrix and σ^2_g the additive genetic variance. Analogous toother methods, the marker effects can be obtained using $\hat{m} = (M'M)^{-1}M'\hat{g}^{**}$.

Principal component correction

The correction via PC method is traditionally used in genomic association (Daetwyler et al. 2012, Riedelsheimer et al. 2012, Price et al. 2006). Similar to EVG, PC correction is based on the inclusion of fixed covariates in the GBLUP model; however, here the covariates are the PCs of the G matrix as follow:

$$y = Xb + \sum_{i=1}^{q} Q_i\beta_i + Zg + e,$$

where q is the number (here 1–50) of the PCs given by $Q = UG$, Q_i represents the columns of the matrix Q, b is the vector of fixed effects with an incidence matrix X, g is the vector of additive random genetic effects of individuals, and e is the residuals vector. These components were fitted as fixed effects because β_i represents the regression coefficients related to the PCs. The marker effects can be obtained using $\hat{m} = (M'M)^{-1}M'\hat{g}$.

Methods comparison

Four scenarios were simulated ten times, with nine of these replications considered as training populations and the remaining replication as the validation population. The EVG, PC, and BLUP methods were compared to the DP method. BLUP and the PC correspond to the population structure correction methods most used in genomic prediction.

The validation population was used to evaluate the agreement between the predicted and parametric breeding values based on Mendelian segregation (i.e., phenotypes without family effects) through four measurements: i) accuracy, given by the correlation between the predicted and the parametric breeding values (simulated); ii) estimation bias, given by the regression coefficient obtained by regression between the predicted and the parametric breeding values with 1 being the ideal value (i.e., absence of bias); iii) molecular heritability given by $h^2 = \frac{\sigma^2_{gM}}{\sigma^2_{gM} + \sigma^2_e}$, where σ^2_{gM} is the additive genomic variance and σ^2_e is the residual variance; iv) agreement between the parametric ordering of the individuals and the ordering obtained for each evaluated method (the method with the highest agreement should be given preference when the objective is individual ranking). Accuracy, bias, heritability, and concordance of ranking of individuals were obtained for each simulation replication, and the reported results were average values.

Computational tools

The mixed model analysis based on the relationship matrix via pedigree was conducted using ASREML® (Gilmour et al. 2009) and the computational routines of the correction methods were implemented in the R software (R Development Core Team 2015) using the *pedigreemm* (BLUP), *rrBLUP* (EVG and PC corrections), *base* (eigenvectors), and *stats* (principal components) packages and the *pedigreemm*, *mixed.solve*, *eigen*, and *princomp* functions, respectively.

RESULTS AND DISCUSSION

Accuracy

The average accuracy of the Mendelian segregation effect estimates is shown in Figure 1. The curves obtained with the EVG method (Figure 1A) show that the accuracy reaches a plateau when 20 eigenvectors are included in the model, whereas a smooth linear decrease is observed when accuracy is calculated using the PC method (Figure 1B). Although, the correction via PC presented higher accuracy than EVG did, the number of eigenvectors is in accordance with that obtained by Janss et al. (2012), who demonstrated that the number of eigenvectors typically used in these analyses is 10 or 20 in differently simulated stratified populations.

According to Daetwyler et al. (2012), the curve representing accuracy against the number of eigenvectors (Figure 1A) reaches a plateau only when LD takes place. Therefore, the curve in Figure 1B, which was obtained using the PC correction method and shows a linear decrease, possibly captured genetic information beyond LD, such as co-segregation (Azevedo et al. 2015).

The average values of the molecular heritability estimates, accuracy, and bias obtained by performing correction via EVG (i.e., 20 eigenvectors), PC (with the inclusion of 1 principal component), BLUP, and DP are shown in Table 1. For accuracy, the correction via EVG (0.62, 0.71, 0.62, and 0.70, for Scenarios 1, 2, 3, and 4, respectively) and PC (0.61, 0.75, 0.60,

and 0.70 for Scenarios 1, 2, 3, and 4, respectively) yielded comparable values for all scenarios and outperformed the other methods. Parametric accuracy values were obtained considering effective population size, chromosome size (in Morgans), number of loci, and heritability of each trait according to Grattapaglia and Resende (2011). Interestingly, the accuracy of the interval estimates obtained by correction via EVG and PC included the parametric value. On the other hand, the accuracy interval estimates obtained by correction via DP did not include the parametric value, whereas the interval estimates obtained by correction via BLUP contained the parametric value for all scenarios, with the exception of Scenario 3. The ranges of the confidence intervals for predictive ability values obtained by correction via EVG and PC were also similar, except for Scenario 4, for which the correction interval via PC had considerably high amplitude compared to that obtained via EVG.

Bias

The bias values estimated by correction via EVG and PC were bigger than 1 (Table 1), whereas by correction via DP they were smaller than 1 (Table 1), indicating that the predicted genetic values were underestimated and overestimated, respectively (Resende et al. 2012). Despite the accuracy obtained by correction via BLUP was higher than obtained via DP, the bias obtained by correction via BLUP was bigger than10, which indicates considerable underestimation of the predicted values.

Molecular heritability

Estimates of heritability in all scenarios were compared with parametric heritability values (0.12 for Scenarios 1 and

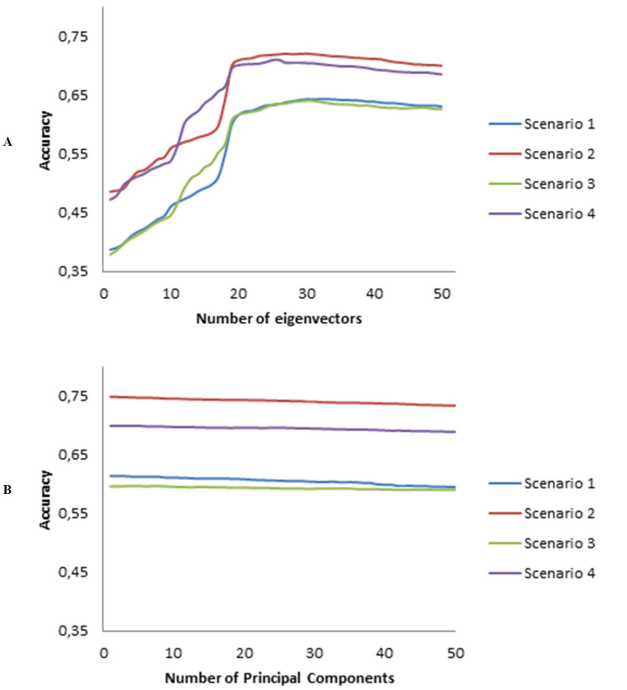

Figure 1. Accuracy involving of predicted and parametric genetic values and the parametric values of the effects of Mendelian segregation for each scenario different scenarios obtained via the covariance analysis via eigenvector correction method (A) and via the principal component correction method (B). The scenarios are defined as follows: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effect genes and heritability 0.20; Scenario 4, trait controlled by small and major effect genes and heritability 0.32.

Table 1. Molecular heritability (h^2), accuracy ($r_{g.\hat{g}}$), and bias estimated for each scenario using various population structure correction methods

	Scenarios	Parametric	BLUP[1]	EVG[2]	PC[3]	DP[4]
h^2	Scenario1	0.12	$0.14_{[0.02;0.26]}$	$0.16_{[0.07;0.24]}$	$0.14_{[0.09;0.20]}$	$0.13_{[0.06;0.20]}$
	Scenario2	0.22	$0.23_{[0.16;0.29]}$	$0.30_{[0.21;0.39]}$	$0.29_{[0.21;0.37]}$	$0.23_{[0.16;0.29]}$
	Scenario3	0.12	$0.11_{[0.08;0.15]}$	$0.14_{[0.02;0.26]}$	$0.13_{[0.07;0.19]}$	$0.11_{[0.08;0.15]}$
	Scenario4	0.22	$0.19_{[0.12;0.26]}$	$0.27_{[0.16;0.38]}$	$0.23_{[0.09;0.36]}$	$0.19_{[0.12;0.26]}$
$r_{g.\hat{g}}$	Scenario1	0.61	$0.38_{[0.06;0.69]}$	$0.62_{[0.54;0.69]}$	$0.61_{[0.55;0.68]}$	$0.31_{[0.26;0.37]}$
	Scenario2	0.66	$0.58_{[0.15;1.00]}$	$0.71_{[0.65;0.77]}$	$0.75_{[0.69;0.81]}$	$0.40_{[0.35;0.46]}$
	Scenario3	0.59	$0.29_{[0.22;0.36]}$	$0.62_{[0.52;0.72]}$	$0.60_{[0.52;0.68]}$	$0.31_{[0.24;0.38]}$
	Scenario4	0.64	$0.60_{[0.22;0.97]}$	$0.70_{[0.64;0.77]}$	$0.70_{[0.57;0.83]}$	$0.38_{[0.31;0.46]}$
Bias	Scenario1	-	>10	$2.79_{[1.86;3.74]}$	$1.66_{[1.46;1.85]}$	$0.92_{[0.43;1.41]}$
	Scenario2	-	>10	$1.87_{[1.57;2.18]}$	$1.41_{[1.30;1.51]}$	$0.77_{[0.67;0.88]}$
	Scenario3	-	>10	$3.38_{[0.61;6.14]}$	$1.67_{[1.55;1.79]}$	$0.94_{[0.61;1.27]}$
	Scenario4	-	>10	$2.11_{[1.62;2.60]}$	$1.47_{[1.31;1.64]}$	$0.82_{[0.66;0.96]}$

[1] Correction for population structure via reduced best linear unbiased prediction (BLUP) computational model; [2] Correction via inclusion of eigenvectors; [3] Correction via inclusion of principal components; [4] Correction via deregressed phenotype (Garrick et al. 2009). The scenarios were defined as follows: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effect genes and heritability 0.20; Scenario 4, trait controlled by small and major effect genes and heritability 0.32.

3 and 0.22 for Scenarios 2 and 4). In Scenarios 1 and 4, the BLUP, PC, and DP estimates were similar to the parametric heritability, whereas the EVG method yielded slightly overestimated results (Table 1). In Scenario 2, the BLUP and DP estimates were similar to the parametric heritability, whereas the EVG and PC estimates were overestimated. In Scenario3, all estimates were similar to the parametric heritability. However, the confidence intervals associated with the four methods all contained the value of the parametric heritability.

Since PCs are estimated based on the G matrix, Janss et al. (2012) and de los Campos and Sorensen (2014) reported that inclusion of PCs as fixed effects in genomic models would carry doubled information from G, which could lead to errors in the variance components and heritability estimation. Price et al. (2010) discussed and advocated the use of the principal components as covariates of fixed effects because the effect of population structure, seen as a function of genetic ancestry, is the same for all the samples. In addition, according to Price et al. (2010), if PCs are adjusted as random effects, spurious associations can be detected, especially those of unusually differentiated markers. In addition, according to Tucker et al. (2014), the restricted maximum likelihood (REML) approach, which consists in projecting the G matrix into a subspace orthogonal to the PCs, effectively removes any PC effect on the variance components and heritability estimation. This agrees with the results reported in this study, in which the heritability values estimated by PC correction are similar to the parametric values in almost all the considered simulation scenarios.

Agreement of ranking

The results obtained by parametric ranking of individuals and the ranking obtained by each of the methods analyzed are shown in Table 2. Although the predicted genetic values found by correction via EVG are more biased than those found via PC, the ranking of this method is considerably more efficient, especially in Scenario 4. In comparison with the other methods, correction via EVG is preferable for individual ranking, especially when the number of individuals is increased. The results obtained by correction via BLUP and DR differ when the number of selected individuals is reduced, but when this number increases, the results are similar.

The main advantage of population structure correction for prediction, independently of the method, is that the estimated marker effects are valid for prediction in various generations, including future generations. The re-estimation of new effects of markers for each generation would involve the development of a new model and heavy time investment, such as for collection of genotypic and phenotypic data and genealogy of the new generation. However, one advantage of correction via EVG and PC over other correction methods is the non-assumption of prior information such as pedigree; in fact, this information is not always available and is often associated with clerical errors. Moreover, correction using eigenvectors or components is associated with a relatively simple theory compared to the DP method, and is performed in one single step, which requires far less computational time than the other methods.

The population structure directly affects association studies and genomic prediction (Guo et al. 2014). Moreover, it is especially important for long-term genomic prediction considering future generations, which requires only

Table 2. Average results for agreement of ranking in each scenario

Scenarios	Top	BLUP[1] (%)	EVG[2] (%)	PC[3] (%)	DP[4] (%)
Scenario1	5	7	16	9	2
	10	09	28	27	7
	50	22	30	30	21
	100	24	38	34	25
	200	39	53	48	36
	250	43	56	52	40
	500	63	72	69	61
Scenario 2	5	13	18	18	09
	10	18	32	23	11
	50	31	35	32	22
	100	38	45	40	30
	200	48	55	48	39
	250	53	59	53	44
	500	70	75	70	63
Scenario 3	5	00	16	13	09
	10	04	19	16	08
	50	11	31	26	14
	100	19	39	30	19
	200	31	50	44	30
	250	36	53	49	36
	500	60	71	68	61
Scenario 4	5	22	20	09	02
	10	21	24	18	11
	50	29	36	26	16
	100	36	44	34	20
	200	47	54	45	31
	250	53	58	51	37
	500	70	75	69	62

[1] Correction for population structure via reduced best linear unbiased prediction (BLUP) computational model; [2] Correction via inclusion of eigenvectors; [3] Correction via inclusion of principal components; [4] Correction via deregressed phenotype (Garrick et al. 2009). The scenarios were defined as follows: Scenario 1, trait controlled by genes with small effects and heritability 0.22; Scenario 2, trait controlled by genes with small effects and heritability 0.37; Scenario 3, trait controlled by small and major effect genes and heritability 0.20; Scenario 4, trait controlled by small and major effect genes and heritability 0.32.

LD information between markers and QTLs (Resende et al. 2012). According to Lehermeier et al. (2015), population structure is an important effect that should be considered in analysis, especially in plant breeding programs, for which the substructure is occasioned between breeding groups or is a consequence of artificial selection and genetic drift.

In this context, population structure effects can be seen as extra-effects affecting "pure" LD information. Thus, corrections for population structure by removing potential population structure effects are required to obtain reliable predictions in future generations (Resende et al. 2012, Albrecht et al. 2014, Guo et al. 2014).

In conclusion, among the methods analyzed in this study, the correction via EVG and PC provided the most accurate genomic breeding estimates values. The correction via BLUP considerably overestimated the genomic breeding estimates values, correction via EVG and PC moderately overestimated, and correction via DP moderately underestimated. The correction via BLUP and DP presented heritability estimates similar to the parametric heritability values. The correction via PC and EVG slightly overestimated the estimates in Scenarios 1 and 3, respectively. The individuals ranking obtained by correction via EVG proved to be considerably more efficient than that of other methods.

REFERENCES

Albrecht T, Auinger HJ, Wimmer V, Ogutu JO, Knaak C, Ouzunova M, Piepho HP and Schön CC (2014) Genome-based prediction of maize hybrid performance across genetic groups, testers, locations, and years. **Theoretical and Applied Genetics 127**: 1375-86.

Azevedo CF, Resende MDV, Silva FF, Viana JMS, Valente MS, Resende MFR and Munoz P (2015) Ridge, Lasso and Bayesian Additive-dominance genomic models. **BMC Genetics 16**: 1: 13.

Calus MPL, Meuwissen THE, Roos APW and Veerkamp RF (2008) Accuracy of genomic selection using different methods to define haplotypes. **Genetics 178**: 553-561.

Carvalho ADF, Neto RF and Geraldi IO (2008) Estimation and prediction of parameters and breeding values in soybean using REML/BLUP and Least Squares. **Crop Breeding and Applied Biotechnology 8**: 219-224.

Daetwyler HD, Kemper KE, Van Der Werf JHJ and Hayes BJ (2012) Components of the accuracy of genomic prediction in a multi-breed sheep population. **Journal of Animal Science 90**: 3375-3384.

de los Campos G and Sorensen D (2014) On the genomic analysis of data from structured populations. **Journal of Animal Breeding and Genetics 131**: 163-164.

Garrick DJ, Taylor JF and Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution 1**: 41:55.

Gilmour AR, Gogel BJ, Cullis BR and Thompson R (2009) **ASReml user guide**. Release 3.0. Available at <https://www.vsni.co.uk/resources/documentation/asreml-user-guide>. Accessed in Sept, 2014.

Goddard ME, Hayes BJ and Meuwissen THE (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. **Journal Animal Breeding and Genetics 128**: 409-421.

Grattapaglia D and Resende MDV (2011) Genomic selection in forest tree breeding. **Tree Genetics and Genomes 7**: 241-255.

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D and Gay G (2014) The impact of population structure on genomic prediction in stratified populations. **Theoretical and Applied Genetics 127**: 749-62.

Janss L, de los Campos G, Sheehan N and Sorensen D (2012) Inferences from genomic models in stratified populations. **Genetics 192**: 693-704.

Lehermeier C, Schön CC and de los Campos G (2015) Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. **Genetics 201**: 323-337.

Oliveira RA, Daros E, Resende MDV, Bespalhok-Filho JC, Zambon JLC, Souza TR and Lucius ASF (2011) Procedimento Blupis e seleção massal em cana-de-açúcar. **Bragantia 70**: 796-800.

Patterson NJ, Price AL and Reich D (2006) Population structure and eigenanalysis. **Plos Genetics 2**: 2074-2093.

Piccoli ML, Braccini J, Cardoso FF, Sargolzaei M, Larmer SG and Schenkel FS (2014) Accuracy of genome-wide imputation in Braford and Hereford beef cattle. **BMC Genetics 15**: 1:15.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. **Nature Genetics 38**: 904-909.

Price AL, Zaitlen NA, Reich D and Patterson N (2010) New approaches to population stratification in genome-wide association studies. **Nature Genetics 11**: 459-463.

R Development Core Team (2015) **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna. Available at <http://www.R-project.org/>. Accessed in June, 2015.

Resende MDV (2002) **Genética biométrica e estatística no melhoramento de plantas perenes**. Embrapa Informação Tecnológica, Brasília, 975p.

Resende MDV and Barbosa MHP (2006) Selection via simulated Blup base on family genotypic effects in sugarcane. **Pesquisa Agropecuária Brasileira 41**: 421-429.

Resende MDV, Silva FF, Lopes OS and Azevedo CF (2012) **Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada e estatística espacial**. Available at <http://www.det.ufv.br/ppestbio / corpo_docente.php>. Accessed in Jan, 2014.

Ribeiro AO, Bearzoti E and Sáfadi T (2005) QTL mapping of Poisson traits: a simulation study. **Crop Breeding and Applied Biotechnology 5**: 310-317.

Riedelsheimer C, Technow F and Melchinger AE (2012) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. **BMC Genomics 13**: 1-9.

Silva FL, Barbosa MHP, Resende MDV, Peternelli LA and Pedrozo CA (2015) Efficiency of selection within sugarcane families via simulated individual BLUP. **Crop Breeding and Applied Biotechnology 15**: 1-9.

Tucker G, Price AL and Berger B (2014) Improving the power of gwas and avoiding confounding from population stratification with pc-select. **Genetics 197**: 1045-1049.

Viana JMS (2004) Quantitative genetics theory for non-inbred populations in linkage disequilibrium. **Genetics and Molecular Biology 27**: 594-601.

Viana JMS (2013) Real breeding. UFV, Viçosa

Yang J, Lee SH, Goddard ME and Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. **The American Journal of Human Genetics 88**: 76-82.