PAULO MAFRA DE ALMEIDA COSTA

PREDICTION OF BREEDING VALUES IN SUGARCANE USING PEDIGREE AND GENOMIC INFORMATION

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2015

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T

C837p
2015

Costa, Paulo Mafra de Almeida, 1985-
Prediction breeding values in sugarcane using pedigree and genomic information / Paulo Mafra de Almeida Costa. - Viçosa, MG, 2015.
v, 25f. : il. ; 29 cm.

Inclui apêndice.
Orientador : Luiz Alexandre Peternelli.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.21-25.

1. Cana-de-açúcar - Melhoramento genético - Métodos estatísticos. 2. Cana-de-açúcar - Seleção. 3. Genética vegetal. 4. Variabilidade (Genética). 5. Marcadores moleculares. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 633.61

PAULO MAFRA DE ALMEIDA COSTA

PREDICTION OF BREEDING VALUES IN SUGARCANE USING PEDIGREE AND GENOMIC INFORMATION

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 17 de dezembro de 2015.

Bruno Portela Brasileiro

Sebastião Martins Filho

Marcos Deon Vilela de Resende
(Coorientador)

Volmir Kist

Luiz Alexandre Peternelli
(Orientador)

AGRADECIMENTOS

A Deus por me iluminar na luta pelos meus sonhos com saúde e perseverança.

Aos meus pais, Flávio e Eliana, pelos sacrifícios e ensinamentos e por me educarem com muito amor e sabedoria. A todos meus familiares por sempre torcerem pelo meu sucesso.

Ao meu irmão Flávio pela amizade e pelo apoio incondicionais.

À Solimar, pela doçura e companheirismo nesses anos de convivência.

A todos os amigos que fiz em Viçosa/MG e Weslaco/TX, os quais contribuíram em minha formação, dividindo momentos de alegria e também de dificuldades.

Ao orientador e amigo Luiz Alexandre Peternelli, pelo apoio e compreensão durante o doutorado, pelos ensinamentos e pelo exemplo de profissional engajado e professor comprometido. Ao Dr. Marcos Deon Vilela de Resende pela coorientação e disponibilidade em ajudar na execução das análises; ao Dr. Márcio Henrique Pereira Barbosa pela coorientação e apoio na execução dos trabalhos, ao Dr. Bruno Portela Brasileiro pela amizade e parceria na execução dos trabalhos, ao Dr. Sebastião Martins Filho pela participação e contribuição na banca de defesa; ao Dr. Volmir Kist pela amizade e pela participação na banca, ao Dr. Fabyano Fonseca e Silva pela ajuda na execução das análises.

A todos os professores e funcionários da UFV, pelos ensinamentos transmitidos, pelo apoio e incentivo em minha caminhada acadêmica, em especial Prof. Cosme Damião Cruz, Sérgio Yoshimitsu Motoike, José Marcelo Soriano Viana, Ana Carolina Campana Nascimento, Felipe Lopes da Silva e Telma Fallieri. Ao Programa de Melhoramento Genético de Cana-de-Açúcar/RIDESA, em especial à equipe do CECA, pelo apoio na execução dos experimentos. Aos amigos da Genética e Melhoramento pelo convívio e motivação À Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Genética e Melhoramento pela oportunidade de realizar meus estudos. Aos cidadãos brasileiros que pagam seus impostos, possibilitando a concessão de bolsa de estudos e financiamento de projetos pelas agências de fomento. Gostaria de agradecer a todos que contribuíram para a realização deste trabalho, direta ou indiretamente, bem como a todos que torcem pelo meu sucesso em busca dos meus sonhos.

SUMÁRIO

RESUMO

COSTA, Paulo Mafra de Almeida, D.Sc., Universidade Federal de Viçosa, dezembro de 2015. **Predição de valores genéticos em cana-de-açúcar usando informação de pedigree e genômica.** Orientador: Luiz Alexandre Peternelli. Coorientadores: Marcos Deon Vilela de Resende e Márcio Henrique Pereira Barbosa.

Aplicações recentes da predição genômica na área vegetal têm incentivado o seu uso em melhoramento de plantas. O desenvolvimento de ferramentas genômicas para acelerar o melhoramento de cana-de-açúcar está atrasado em comparação com outras grandes culturas, portanto, estudos empíricos devem ser realizados para avaliar a utilidade desta abordagem para o melhoramento desta importante cultura. Os objetivos desse trabalho foram: i) avaliar a acurácia da predição de quatro caracteres quantitativos de cana-de-açúcar usando informação de marcadores SNPs e ii) comparar acurácias entre predições usando pedigree e informação genômica. Valores genéticos foram preditos em uma população de melhoramento da fase T2 de 514 indivíduos genotipados com 37.024 marcadores SNP. Cinco modelos preditivos foram avaliados: Genomic BLUP (GBLUP), Bayesian LASSO (BL), Bayes A (BA), Bayes B (BB) e Bayesian Ridge Regression (BRR). As acurácias dos métodos foram avaliadas através da correlação entre valores genéticos preditos e observados por meio de validação cruzada. Os métodos apresentaram valores de acurácia muito semelhantes. No entanto, houve diferenças marcantes nas acurácias obtidas entre características. A maior acurácia foi obtida para fibra pelo método BRR (0,57) e a menor foi obtida para toneladas de pol por hectare pelo método Bayes B (0,07). Na comparação da predição utilizando pedigree e informação genômica exibiu valores mais elevados de correlação, bem como desvio padrão inferior, exceto para toneladas de cana por hectare e toneladas de pol por hectare. A informação genômica explicou maior proporção da variância genética em comparação com o pedigree. Acurácias satisfatórias foram obtidos utilizando informação genômica, especialmente para percentual de pol em cana e porcentagem de fibra em bagaço. Assim, sua utilização pode ser um abordagem eficaz para a melhoria das características agronômicas desejáveis em uma cultura poliplóide complexa.

ABSTRACT

COSTA, Paulo Mafra de Almeida, D.Sc., Universidade Federal de Viçosa, December 2015. **Prediction of breeding values in sugarcane using pedigree and genomic information.** Adviser: Luiz Alexandre Peternelli. Co-advisers: Marcos Deon Vilela de Resende and Márcio Henrique Pereira Barbosa.

Recent applications of genomic prediction in plant science have encouraged its use in plant breeding. The development of genomic tools for speed up sugarcane breeding has been delayed compared to other major crops, therefore, empirical studies must be conducted to evaluate the usefulness of this approach on this important crop. The objectives of our study were: i) to assess the accuracy of prediction of four quantitative traits using genomic information of SNP markers in a commercial sugarcane breeding population; ii) to compare the accuracy between predictions using pedigree and genomic information. Genetic values were predicted in a second phase trial population of 514 individuals genotyped with 37,024 SNP markers. Five statistical predictive models were evaluated: Genomic BLUP (GBLUP), Bayesian LASSO (BL), Bayes A (BA), Bayes B (BB) and Bayesian Ridge Regression (BRR). Accuracies of the methods were assessed through the correlation between observed and predicted genetic values in a 10-fold cross validation. The methods exhibited very similar accuracy values regarding the trait. Nevertheless, there were marked differences among traits. The highest accuracy was obtained for FB by the BRR method (0.57) and the lowest was obtained for TPH by Bayes B method (0.07). Two models (pedigree – P and pedigree + genomic – P+G) were fitted and used to predict the traits in order to compare the prediction accuracy using pedigree and genomic information. Overall, P+G exhibited higher correlation values, as well as lower standard deviation, except for the traits TSH and TPH. Genomic information explained higher proportion of the genetic variance in comparison to the pedigree. Satisfactory accuracies were obtained by using genomic information, especially for pol percentage in sugarcane and fiber percentage in bagasse. Thus, the use of genomic information could be more efficient per unit of time for improvement of desirable agronomic traits in a complex polyploid crop.

Introduction

Sugarcane (*Saccharum spp.*) is one of the crops of higher economic importance worldwide, being cultivated in more than 100 tropical and subtropical countries for sugar, ethanol and, ultimately, for bioenergy purposes (Matsuoka et al. 2014; da Silveira et al. 2015). Given the growing existing demand for its main products, increases in crop area and yield must be recorded annually. Thus, the search for varieties highly productive and adapted to the different environments of cultivation has been intensified.

Improvement of desirable agronomic traits requires large experiments that last for several crop cycles. About 10 to 12 years of field experiments are needed to release superior sugarcane cultivars (Loureiro et al. 2011). The development of genomic tools (e.g., marker assisted selection) for speed up sugarcane breeding has been delayed compared to other major crops such as maize, rice, soybean, wheat, etc. This is in part due to the complexity of the sugarcane genome, which makes marker-trait association and QTL studies challenging (D'Hont et al. 2008; Henry and Kole 2010; Hotta et al. 2010; Souza et al. 2011; Dal-Bianco et al. 2012).

Currently, sugarcane breeding relies on phenotypic selection by means of mass and/or family selection strategies (Stringer et al. 2011; Barbosa et al. 2012). Superior clones are identified and selected based on its predicted genotypic values. The evaluation of potential clones to be used as parents in crosses is based on its predicted breeding values (BV); therefore, an accurate prediction of them is crucial to increase genetic gain in any breeding program (Falconer and Mackay 1996). To this end, Restricted Maximum Likelihood (REML) and Best Linear Unbiased Prediction (BLUP) are optimum methodologies widely adopted for these purposes (Henderson 1975; Piepho et al. 2008).

In a traditional BLUP procedure, breeders can use progeny data and relatedness among individuals to predict breeding values, which is done by including pedigree information. The inclusion of pedigree information is performed via the numerator relationship matrix (A) computed from the coefficient of relationship (Henderson 1984; Falconer and Mackay 1996; Mrode 2005). It is well known that BV estimation using information from pedigree results in a better predictive accuracy due to the exploitation

of the genetic correlation among relatives (Piepho et al. 2008; Bernardo 2010; Viana et al. 2011; Viana et al. 2012). The higher the genetic correlation of a genotype of interest with related genotypes, the more information can be gained from records of relatives (Piepho et al. 2008). Atkin et al. (2009) showed that adding pedigree information back to the base population improved the estimates of additive variance components and breeding values of sugarcane parents.

It should be stressed, however, that the estimation of genetic effects via the coefficient of relationship relies on several assumptions related to the underlying quantitative-genetic theory, which usually are not fully met in plant breeding data (Piepho et al. 2008). The genetic relationships estimated from pedigree are approximations, i.e., an expected value determined by the averaged alleles shared by the relatives that traced back to a common ancestor (Falconer and Mackay 1996). On the other hand, genetic relationships based on many DNA markers genotyped across the genome are more precise because high-density genotyping identifies genes identical in state instead of those identical by descent (Lynch and Walsh 1998; VanRaden 2008). Therefore, the frequency of alleles shared by individuals can be used to construct realized genomic relationships (G matrix, VanRaden 2008) aiming at more accurate predictions of the genetic values of candidate clones.

Besides the use of genomic information to substitute pedigree in genetic evaluations, recent applications of genomic prediction (first proposed by Meuwissen et al. 2001) in several crops have encouraged its use in plant breeding. Successfully applied in animal breeding (Meuwissen et al. 2013; Daetwyler et al. 2013; de Los Campos et al. 2013), empirical studies have been conducted to demonstrate its feasibility in plant science with encouraging results (Lin et al. 2014).

Concerning to sugarcane, the only result of application of genomic prediction reported up to date is from Gouy et al. (2013). The authors found no major differences in prediction accuracies among four methods evaluated (Bayesian LASSO, Ridge Regression, Reproducing Kernel Hilbert Space and Partial Least Square Regression). Accuracies ranged from 0.11 to 0.62 within a panel of 167 accessions representing a core sampling of elite germplasm. Despite the number of DArT markers used was not

sufficient to densely cover the large sugarcane genome, results demonstrate the potential value of genomic prediction for sugarcane breeding.

Accordingly, the objectives of our study were: i) to assess the accuracy of prediction of four quantitative traits using genomic information of SNP markers in a commercial sugarcane breeding population; ii) to compare the accuracy between predictions using pedigree and genomic information.

Materials and Methods

Plant material

The population used in this study consisted of 514 clones from 98 half-sib families originating from elite clone parents and commercial varieties. The genitors were chosen among clones created through intercrossing and commonly used in crosses in the RIDESA/ UFV Sugarcane Breeding Program (Barbosa et al. 2012).

Crossings were performed at the Flowering and Crossings Station, Murici, Alagoas, Brazil (09°13' S, 35°50' W, 450-m altitude). Seedlings obtained from each family were transplanted in 2010 at the Centro de Pesquisa e Melhoramento da Cana-de-Açúcar of the Uciversidade Federal de Viçosa, Oratórios, Micas Gerais, Brazil (20°25' S, 42°48' W, 494-m altitude). The seedlings were conducted in the first phase (T1) trials (Loureiro et al. 2011; Barbosa and da Silveira 2012) in 2011 (plant cane) and 2012 (first-ratoon). The population used in this study was originated from seedlings selected in T1 and advanced to the second phase (T2) trial of the breeding program.

Phenotypic data

Phenotypic measurements were obtained from T2. The trial was conducted in an augmented block design (Federer 1961; Baffa et al. 2014). The clones (unreplicated) and two common reference cultivars (replicated once each) were arranged in 49 augmented blocks. The reference cultivars were RB867515 and SP80-1842, which are cultivated in large areas and commonly used in breeding experiments.

The trial was established in July 2012 in an experimental field located at the Centro de Pesquisa e Melhoramento da Cana-de-Açúcar, Oratórios, Minas Gerais, Brazil (20°25' S, 42°48' W, 494-m altitude) on a Rhodic Eutrudox soil. The experimental plots

consisted of 5-m rows with a spacing of 1.4 m between rows. The clones were evaluated in plant cane in July 2013 (after 12 months of growth).

The following traits were evaluated: apparent percentage of sucrose in sugarcane (PC), tonnes of stalks per hectare (TSH), tonnes of pol per hectare (TPH), and percentage of fiber in sugarcane bagasse (FB).

Ten stalks taken randomly from each row were used for estimating field productivity and juice quality parameters. TSH was estimated from the total number of stalks per row and the wet weight of 10 stalks determined with a dynamometer; FB was determined on wet basis from a 500-g sample of shredded stalk (Legendre 1992), as follows:

$$FB(\%)=(0.1417 \times WM - 7.8333) \quad (1)$$

where WM is the wet mass of the sample removed from the hydraulic press. The apparent percentage of sucrose in juice (POL) was measured by polarimetric determination after juice extraction from 500-g samples crushed in a hydraulic press (Schneider 1979) and used to derive the apparent percentage of sucrose in sugarcane (PC), according to the following expression (Baffa et al. 2014):

$$PC(\%)=POL\times(1-0.01\times FB)\times(0.9961-0.0041\times FB) \quad (2)$$

where POL and FB are measurements of apparent percentage of sucrose in juice and percentage of fiber in sugarcane bagasse, respectively. The trait TPH, expressed as a percentage of apparent sucrose on a fresh weight basis, was estimated as follows:

$$TPH = (TSH \times PC) / 100 \quad (3)$$

Genotypic data

Genomic DNA was extract using DNeasy Plant Mini Kit from Qiagen® following manufacturer's guidelines. A minimum of 3 ug per sample was sent for genotyping to the company RAPiD Genomics (http://www.rapid-genomics.com), Florida, USA.

Samples were genotyped for SNPs markers by means of a technique based on the Capture Seq® technology. Briefly, a set of 50,000 unique sequences was identified

from: i) existing expressed sequence tags (ESTs) from public sugarcane cDNA libraries and ii) whole-shotgun genome sequences available publically, consistently distributed in the genome and assuming synteny to the sorghum genome. Biotinylated 120-mer probes that complement a segment of each of the 50,000 target regions of the sugarcane genome were synthesized and were utilized to capture sequences at each target locus. After the next-generation sequencing, data were aligned to the reference used to design the probes and SNPs were identified and genotyped based on the alignment results and information about the population (Neves et al. 2013).

The raw data consisted of raw sequences (reads) and data files filtered by the RAPiD Genomics' quality parameters. The filters were applied to the raw sequences and were based on alignment quality, base calling quality and coverage depth. The following quality parameters were adopted in order to generate a raw SNP markers matrix:

i) A minimum phred-scaled quality score of 20. The phred score measures the assertion made in non-reference (alternate) alleles calling, according to the expression:

$$QUAL=-10\log_{10}\times \text{prob (call alternate allele wrongly)} \qquad (4)$$

Thus, a quality of 20 means that SNPs with a probability of miscalling higher than 1% were eliminated.

ii) A minimum read depth of 5. It indicates how many times a fragment was sequenced. Thus, reads with depth less than 5 were eliminated, decreasing the chance of having false polymorphism as SNPs.

iii) A minimum of missing data of 20%. It indicates that SNPs with missing data higher than 20% were eliminated.

iv) A minimum MAF (minor allele frequency) of 3%. It indicates that SNPs with MAF less than 3% were eliminated, thus excluding the non-informative markers.

A total of 89,288 SNPs were obtained. The raw SNP markers matrix X (89,288 SNPs and 514 individuals) was coded as follows:

$$\begin{cases} 0;\ mm, \text{ denoting homozygozity for the reference allele} \\ 1;\ Mm, \text{ denoting heterozygozity, with one reference and one alternative allele} \\ 2;\ MM, \text{ denoting homozygozity for the alternative allele} \end{cases}$$

After coding, the X matrix was submitted to standard quality control parameters in genomic prediction studies, performed by means of *HapEstXXR* package (Knueppel and Rohde 2015). The functions implemented in the package perform calculations of MAFs, call rate and asymptotic chi-square test. The number of missing genotypes divided by sample size defines the call rate. Call rate ≥ 0.95 and MAF ≥ 0.05 were used as quality parameters, therefore, markers with missing data higher than 5% and minor allele frequency lower than 5% were removed. Missing genotypes were imputed from the density x_{ij}=Binomial $\left(3, \hat{p}_j\right)$, where $\hat{p}_j$ is the estimated allele frequency computed from the non-missing genotypes. Loci were tested for deviation of the Hardy-Weinberg equilibrium by the goodness-of-fit chi-square test. Those loci presenting deviation of the equilibrium were removed (Falconer and Mackay 1996; Laurie et al. 2010). The number of SNPs after edition was 37,024.

Statistical analysis

The phenotypic data were analyzed by means of mixed model methodology using the software Selegen-REML/BLUP (Resende 2007). Variance components were estimated by REML and the genotypic values of the clones were predicted by BLUP. The following mixed linear model was used:

$$y = Xr + Zg + Wb + e \quad (5)$$

where *y* is the vector of phenotypic observations for each trait, *r* is the vector of fixed effects (overall mean), *g* is the vector of random genotypic values, *b* is the vector of random effects of blocks, *e* is the vector of residuals ($e \sim N(0, \sigma_e^2)$), and X, Z, and W are incidence matrices relating the observations to the respective model effects.

Analysis of deviance was done to test the significance of the effects of genotypes and blocks. The statistical significance of the effects was tested by means of the likelihood ratio test. It compares two models (the simpler model has fewer parameters than the general model) and is computed as a difference in the deviance for the two models (Casella and Berger 2001). The predicted genotypic values of the clones ($\mu+\hat{g}$) were used to compound the *y* vector (response variable) in the genomic prediction models.

In order to estimate the additive genetic effects of the clones, the following mixed linear model was used (Resende 2007):

$$y = Xr + Zf + Wb + Sc + e \quad (6)$$

where *y* is the vector of phenotypic observations for each trait, *r* is the vector of fixed effects (overall mean), *f* is the vector of random effects of half-sib families, *b* is the vector of random effect of blocks, *c* is the vector of random effects of clones within families, *e* is the vector of residuals ($e \sim N(0, \sigma_e^2)$), and X, Z, W and S are incidence matrices relating the observations to the respective model effects.

Heritability of the additive genetic effects is given by (Resende 2002):

$$h_a^2 = \sqrt{\frac{4 \times \sigma_f^2}{\sigma_y^2}} \quad (7)$$

where σ_f^2 is the genotypic variance between families and σ_y^2 is the individual phenotypic variance.

Genomic prediction methods

In whole-genome prediction, statistical models are fitted to predict phenotypes using dense molecular markers genotyped across the genome, which may be, by principle, in linkage disequilibrium (LD) with quantitative trait loci (QTLs). Thus, genomic estimated breeding values (GEBV) are function of the markers genotyped in the individuals and are obtained by summing up its estimated effects (Meuwissen et al. 2001).

Five statistical methods were evaluated: Genomic BLUP (GBLUP), Bayesian LASSO regression (BL), Bayes A (BA), Bayes B (BB) and Bayesian Ridge Regression (BRR). The models were fitted (see Appendix for R codes) according to the guidelines of the BGLR package for R (Pérez and de los Campos 2014). Prior densities for regression coefficients and hyperparameters were chosen according to Pérez and de los Campos (2014).

GBLUP uses a molecular additive relationship matrix G (VanRaden 2008), as follows:

$$y = I\mu + Zg + e \quad (8)$$

where y is the vector of genotypic values computed from model (4), μ is the overall mean, I is a column vector of 1s, g is the vector of random additive values (GEBV) of individuals ($g \sim N(0, G\sigma_g^2)$), in which σ_g^2 is the additive genetic variance, Z is a diagonal design matrix and e is the vector of residual effects ($e \sim N(0, \sigma_e^2)$). The G matrix was computed from de SNP marker matrix following the guidelines of the BGLR package (Pérez and de los Campos 2014).

For the other methods, GEBV were obtained as follows:

$$GEBV = X\hat{\beta} \quad (9)$$

where $\hat{\beta}$ is the vector of estimates of marker effects and X is the genotyping matrix.

In Bayesian LASSO regression, markers are assumed to have different variances (heterogeneous shrinkage) and $\hat{\beta}$ are obtained by solving the following optimization problem:

$$\hat{\beta}=\min\left\{\sum_{i=1}^{n}(y_i-x_i'\beta)^2+\lambda_L\sum_{j=1}^{p}|\beta_j|\right\}, \quad (10)$$

where y_i is the observed value of individual i, x_i is the i^{th} row of the markers matrix, β is the corresponding vector of regression coefficients, λ_L the regularization parameter and β_j the estimate effect of the j^{th} marker (de Los Campos et al. 2013). The Gibbs sampler was run for 50,000 iterations, with the first 5,000 discarded as the burn-in and a thinning interval of 10.

Other Bayesian approaches (BA and BB) differ about the specifications of the prior variance of the marker-specific regression coefficient as well as the type and extent of shrinkage. The prior in Bayes B assumes the variance of markers to equal zero with probability p, and the complement with probability (1 – p). Bayes B converges to Bayes A as the probability of non null effects of the markers approaches 1.0.

In BRR, all effects are shrunk to a similar extent (homogeneous shrinkage). It assumes a common variance for all markers. The posterior means of marker effects is given by:

$$\hat{\beta}=(X'X+\lambda I)^{-1}X'y \quad (11)$$

where X and I are the genotyping and identity matrices, respectively, *y* is the vector of phenotypes and λ is the ridge parameter.

Due to the common variance assumed for all markers, BRR may not be appropriate if some markers are not linked to QTL, i.e., they are not associated with genetic variance (Meuwissen et al. 2001; de Los Campos et al. 2013). Therefore, methods using different priors of marker effects, such as BL (double-exponential prior), BA (scaled-t) and BB (two component mixture prior with a point of mass at zero and a sclaed-t slab) were evaluated to overcome this limitation.

Cross-validation and accuracy of genomic selection

The accuracies of the models were evaluated by a 10-fold cross-validation scheme. For each data set, the 514 individuals were randomly split into ten subsets of which nine were used as the training set (used to estimate the marker effects) to predict the remaining tenth (validation set). For comparison purposes, the random subsampling was fixed for all models. The process was repeated 10 times, each time with a different set of individuals as the validation partition, until all individuals had their phenotypes predicted (Resende et al. 2012). The predictive ability was estimated as the correlation between the genetic values predicted by the models and the observed genetic values, using the mean Pearson correlation coefficient among folds. The accuracy of the prediction of additive genetic effects was estimated according to the expression (Resende et al. 2012):

$$r_{\hat{a}a} = r_{\hat{y}y}/h_a \quad (12)$$

where $r_{\hat{y}y}$ is the correlation coefficient between observed and predicted genetic values and h_{awf} is the square root of the heritability of the additive effects within families, obtained by the expression (6).

Comparison of the prediction accuracy using pedigree and genomic information

Two models (pedigree – P and pedigree + genomic – P+G) were fitted and used to predict the traits in order to compare the prediction accuracy using pedigree and

genomic information (Pérez and de los Campos 2014). The pedigree-only model (P) was similar to GBLUP, except that it used the A matrix of additive relationship computed from pedigree, instead of G. In the pedigree + genomic BLUP model (P+G) two random effects are included in the model: one representing a regression on pedigree $a \sim N(0, A\sigma_a^2)$, where A is a pedigree-derived numerator relationship matrix, and one representing a linear regression on markers, $g \sim N(0, G\sigma_g^2)$, where G is a marker-derived genomic relationship matrix. The numerator relationship matrix (A) was obtained by means of the algorithm implemented on R by Peternelli et al. (2009) using pedigree information traced back to the grand-parents of each clone (2 generations). Both matrices were standardized to an average diagonal value of (approximately) one for ease of interpretation of estimates of variance components.

Prediction accuracies of the models including genomic *vs* pedigree information were assessed by multiple training–testing partitions, following the guidelines of the BGLR package (Pérez and de los Campos 2014). In this approach, the data set was partitioned into two sets: one used for model training (TRN, used to estimate the marker effects) and one used for testing (TST, used to prediction and model validation). One hundred individuals were randomly assigned to TST set and the remaining (~ 80% of the population) was used for TRN set. For comparison purposes, the random subsampling was fixed for both models. The process was repeated 100 times, each time with a different set of individuals as the testing partition. The predictive ability was estimated as the correlation between the predicted and the observed genetic values, resulting in 100 estimates of the prediction correlation for each of the models fitted. Since correlations for each of the models are paired samples resulted from the same TRN–TST partition, a paired-t-test based on the difference of the correlation coefficients was conduct for testing the null hypotheses that P and P+G have the same prediction accuracy (Pérez and de los Campos 2014).

Results

Phenotypic data

Analysis of deviance indicated statistical significance for all random effects for all traits (Table 1), suggesting that there is genotypic variability exploitable for breeding

of these traits. The population showed considerable phenotypic variation in the traits: PC ranged from 6.50 to 17.00, with an average of 13.80; TSH ranged from 51.54 to 152.23, with an average of 106.76; TPH ranged from 6.81 to 20.66, with an average of 14.67 and FB ranged from 8.39 to 14.64, with an average of 11.08.

Estimates of variance components and genetic parameters for PC, TSH, TPH and FB are presented in Table 2. Narrow-sense heritabilities varied from 0.17 to 0.55. The estimate of h^2_{awf} for PC was higher than those found for TSH, TPH and FB, indicating a high genetic variance and/or low environmental variance for this trait. Coefficients of genetic variation varied from 10.3% to 18.6%. Similar values were found by Baffa et al. (2014) in a sugarcane breeding population originated from 13 half-sib families.

Table 1 – Statistical analysis of the traits apparent percentage of sucrose in cane (PC), tonnes of stalks per hectare (TSH), tonnes of pol per hectare (TPH) and percentage of fiber in bagasse (FB) of the T2 breeding population evaluated at the plot level.

Effect	PC	TSH	TPH	FB
	Deviance[+]			
Genotype	1454.12	5638.05	2606.98	1735.18
Block	1429.69	5656.42	2589.14	1736.66
Error				
Full model	1404.64	5608.77	2581.71	1709.75
	LRT (χ^2)‡			
Genotype	49.48**	29.28**	25.27**	25.43**
Block	25.05**	47.65**	7.43**	26.91**
Error	-	-	-	-
Full model	-	-	-	-
	Coefficients of determination			
Genotype	0.71	0.57	0.58	0.56
Block	0.07	0.11	0.04	0.08
Error	0.22	0.32	0.38	0.36
Full model	-	-	-	-
	Genetic coefficient of variation, %			
	10.30	18.50	18.63	13.42
Mean	13.788	106.764	14.675	11.076

*Significant at $P \leq 0.05$; ** Significant at $P \leq 0.01$

[+] Deviance of the adjusted model without corresponding effects;

‡Likelihood ratio test: the difference in the deviance of the two models (full model and simpler model), asymptotically distributed as a chi-squared random variable, with degrees of freedom equal to the difference in the number of parameters between the two models.

Table 2 - Estimates of variance components and genetic parameters for apparent percentage of sucrose in sugarcane (PC), tonnes of stalks per hectare (TSH), tonnes of pol per hectare (TPH) and percentage of fiber in sugarcane bagasse (FB) of the T2 breeding population of 514 clones evaluated at the plot level.

Parameters[1]	PC	TSH	TPH	FB
σ_g^2	2.02	389.98	7.48	2.21
σ_b^2	0.19	74.48	0.46	0.31
σ_e^2	0.63	219.64	4.86	1.43
σ_f^2	2.84	684.10	12.80	3.95
h_a^2	0.55	0.17	0.24	0.32
Ac_{gen}	0.86	0.78	0.77	0.77

[1]Genotypic variance (σ_g^2); variance between blocks (σ_b^2), residual variance (σ_e^2); individual phenotypic variance (σ_f^2); narrow-sense heritability of additive effects within families (h_{awf}^2); accuracy of genotypic selection (Ac_{gen}).

Genomic prediction methods

Boxplots of the correlation coefficients between observed and predicted genetic values of five genomic prediction methods are presented in Figure 1. The methods exhibited very similar correlation values regarding the trait. Nevertheless, there were marked differences among traits. Overall, the ability to predict phenotype varied from 0.033 (SD ±0.147) for TPH using Bayes B method to 0.322 (SD ±0.081) for FB using Bayesian Ridge Regression. PC and FB presented the highest mean correlations, with coefficients values varying among methods from 0.304 (SD ±0.07) to 0.310 (SD ±0.07) and 0.317 (SD ±0.08) to 0.322 (SD ±0.08), respectively. TSH and TPH had the lowest mean correlations values, varying from 0.089 to 0.097 and 0.033 to 0.040,

respectively. The traits TSH and TPH exhibited the highest standard deviation of the correlation coefficients (±0.152 to ±0.163 and ±0.139 to ±0.152, respectively).

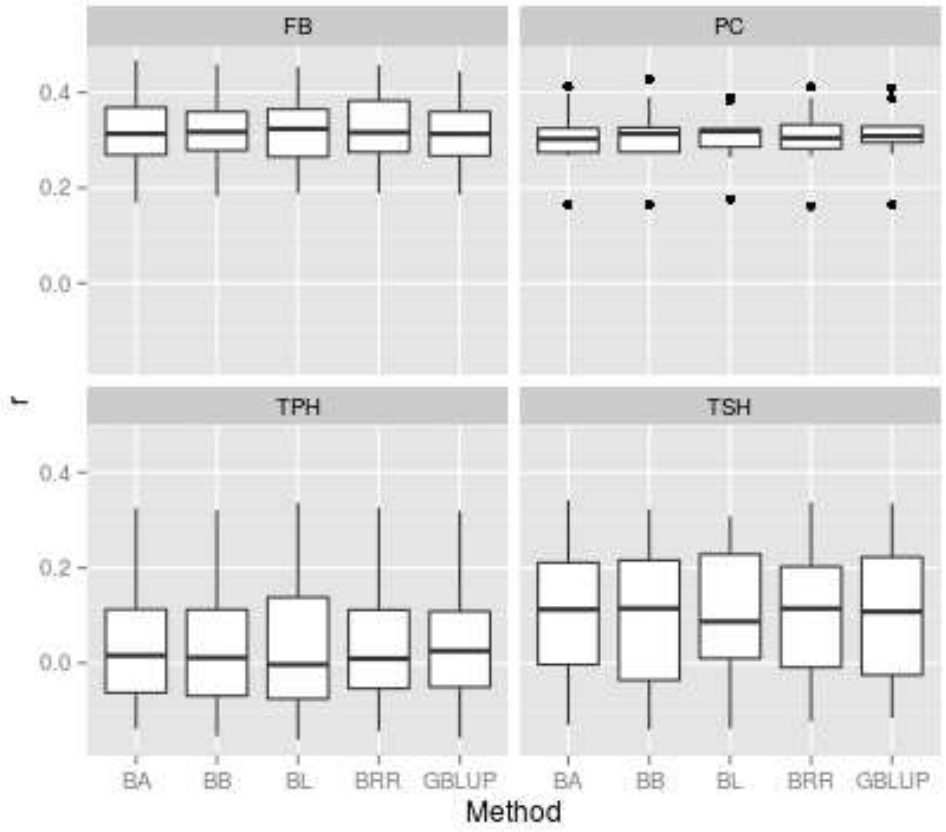

Figure 1 – Boxplots of the correlation coefficients between observed and predicted genetic values of five genomic prediction methods obtained by 10-fold cross validation in a T2 sugarcane breeding population of 514 clones.

Accuracies exhibited similar values among methods as well, ranging from 0.41 to 0.42 for PC, 0.22 to 0.24 for TSH, 0.07 to 0.08 for TPH and 0.56 to 0.57 for FB (Table 3). The highest accuracy was obtained for FB by the BRR method (0.57) and the lowest was obtained for TPH by Bayes B method (0.07). The coefficient of regression (Beta) of GEBV on phenotype of all methods had low magnitude values, varying from 0.610 to 0.615 for PC, 0.014 to 0.016 for TSH, 0.042 to 0.052 for TPH and 0.69 to 0.78 for FB (Table 3).

Table 3 - Accuracy and coefficient of regression of the prediction of additive effects using five genomic prediction methods in 10-fold cross validation conducted in a T2 sugarcane breeding population of 514 clones.

Traits	Statistics	Method				
		GBLUP	BL	BAYES-A	BAYES-B	BRR
PC	Accuracy	0.419	0.415	0.412	0.417	0.415
TSH		0.238	0.232	0.239	0.220	0.228
TPH		0.081	0.071	0.074	0.068	0.072
FB		0.561	0.566	0.561	0.566	0.570
PC	Beta Regression	0.615	0.613	0.610	0.613	0.611
TSH		0.015	0.014	0.015	0.016	0.015
TPH		0.049	0.052	0.042	0.045	0.042
FB		0.710	0.780	0.692	0.763	0.714

[a]Traits: apparent percentage of sucrose in sugarcane (PC), tonnes of stalks per hectare (TSH), tonnes of pol per hectare (TPH) and percentage of fiber in sugarcane bagasse (Fiber);

Comparison of the prediction accuracy using pedigree and genomic information

The mean correlation coefficients between observed and predicted values obtained using P and P+G models are described in Table 4. Overall, P+G exhibited higher correlation values, as well as lower standard deviation, except for the traits TSH and TPH. The correlation values differed statistically by the paired-t test, except for TPH. It indicates that P and P+G models differed on the predictive ability, except for TPH. Figure 2 shows the estimated correlations between observed and predicted values assessed by 100 training-testing partitions in the P and P+G models. The majority of the above the 45º lines indicates that in most partitions the P+G model had higher prediction accuracy than the P model.

Table 4 – Mean correlation coefficients between observed and predicted values for PC, TSH, TPH and FB obtained using P and P+G models from 10-fold cross validation analysis of a T2 sugarcane breeding population of 514 clones.

Traits	Method[a]		
	P	P+G	t^b
PC	0.231 (±0.113)	0.311 (±0.06)	16.71**
TSH	0.004 (±0.203)	0.004 (±0.213)	7.66**
TPH	0.006 (±0.227)	0.004 (±0.196)	-0.18
FB	0.182 (±0.095)	0.305 (±0.084)	14.92**

[a] P: pedigree model; P+G: pedigree + genomic model
[b] Paired-t test
** Significant at $P \leq 0.01$

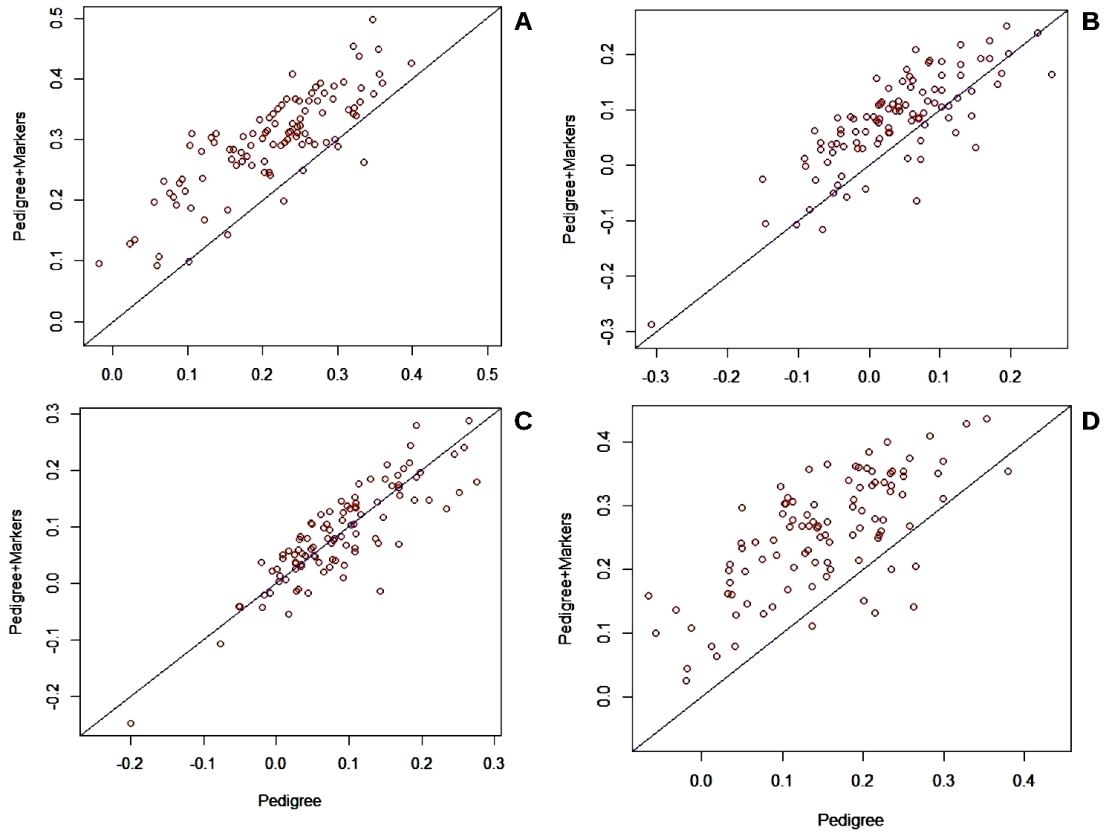

Figure 2 - Estimated correlations between observed and predicted values for: (A) PC, (B) TSH, (C) TPH and (D) FB; using P and P+G models in the multiple training–testing partitions.

Estimates of parameters of the PGBLUP model fitted for comparison of the prediction accuracy using pedigree *vs* genomic information is described in Table 5. Since the phenotypes were standardized to a unit sample variance, the estimated residual variance of 0.35, 0.55, 0.52 and 0.46 for PC, TSH, TPH and FB, respectively, indicates that the P+G model explained approximately 65, 45, 48 and 54% of the phenotypic variance for PC, TSH, TPH and FB, respectively.

Except for TPH, markers explained higher proportion of the genetic variance (r^2_{g*}) in comparison to the pedigree. For TPH, 54% of the genetic variance was explained by the pedigree. The estimated posterior mean of the ratio of the genetic variance relative to the total variance (h^2_m) was 0.66, 0.48, 0.49 and 0.54 for PC, TSH, TPH and FB, respectively.

Table 5 - Posterior means estimates of parameters of the P and P+G models fitted for analysis of four sugarcane traits evaluated in a T2 breeding population.

Parameter[1]		**Trait**			
		PC	**TSH**	**TPH**	**FB**
σ^2_e	Mean	0.351	0.548	0.521	0.460
	SD	0.101	0.114	0.111	0.105
σ^2_{a*}	Mean	0.271	0.229	0.280	0.239
	SD	0.090	0.079	0.096	0.086
σ^2_{g*}	Mean	0.403	0.286	0.232	0.315
	SD	0.122	0.100	0.082	0.092
r^2_{g*}	Mean	0.593	0.552	0.457	0.570
	SD	0.125	0.122	0.124	0.118
r^2_{a*}	Mean	0.407	0.448	0.543	0.430
	SD	0.125	0.122	0.124	0.118
h^2_m	Mean	0.656	0.482	0.493	0.544
	SD	0.099	0.104	0.105	0.101

[1]Residual variance (σ^2_e); variance associated to the pedigree (σ^2_{a*}); variance associated to the markers (σ^2_{g*}); proportion of the genetic variance explained by the markers (r^2_{g*}), proportion of the genetic variance explained by the pedigree (r^2_{a*}); ratio of the genetic variance relative to the total variance (h^2_m)

Discussion

Here we presented the performance of five predictive models for genomic prediction of four agronomic traits in a complex genome crop. The dataset was generated from a breeding population of 514 clones evaluated in the second phase trial of the breeding program. This is the second study assessing the feasibility of genomic

prediction in sugarcane and the first using a commercial breeding population genotyped with SNP markers.

A total of 37,024 SNPs was used to predict phenotypes, a significant increase in the number of markers in comparison with previous study (Gouy et al. 2013). Concerning the type of markers used, it was not possible to draw conclusions about the difference between DArT (dominant) and SNP (codominant) markers in accuracy prediction, since difference exists in population nature and size, traits and marker coverage between both studies. Nevertheless, correlations between observed and predicted genetic values did not differed significantly from those obtained by Gouy et al. (2013). For comparison purposes, the traits PC and FB had mean correlations of 0.31 and 0.32, respectively, in comparison with 0.42 to 0.55 and 0.24 to 0.34 obtained for juice brix and bagasse content, respectively, by those authors.

The values of accuracy found were relatively close to those found in other genomic selection studies conducted with empirical data (Crossa et al. 2010; Resende et al. 2012; Würschum et al. 2013; Crossa et al. 2013b; Gouy et al. 2013; Crossa et al. 2013a), keeping the peculiarities of each crop and experimental conditions. One exception was found for yield traits: the lowest accuracies were found for the TSH (0.22) and TPH (0.07) (Table 3).

As stated by Crossa et al. (2010), larger gains in predictive ability could be obtained using more markers or by improving upon genomic prediction methods. No major differences were found among the five methods evaluated, which is a finding consistent with previous studies (Heslot et al. 2012; Resende et al. 2012; Gouy et al. 2013). Thus, our results suggest that marker coverage seems to be a key factor to obtain more accurate predictions in a large genome crop. Higher marker density would permit stronger linkage disequilibrium between marker loci and QTLs controlling those traits. Moreover, the models fitted are expected to perform better in traits where additive effects play a central role (Crossa et al. 2010), as observed for PC and FB. Considering that yield traits are supposed to be highly influenced by dominant effects (Hogarth et al. 1981; Bastos et al. 2003; Zeni Neto et al. 2013), models accounting for allelic interaction should be further evaluated.

The inclusion of markers (P+G) increased the predictive ability of genetic values in sugarcane, explaining a higher proportion of the genetic variance relative to the pedigree. This is an important finding and suggests that markers may help to capture genetic variance not achieved with pedigree-based predictions. The ratio of the genetic variance relative to the total variance obtained by genomic information was relatively close to the heritability obtained by phenotypic selection, indicating that molecular markers could be used for inference and selection in a sugarcane breeding population.

Low superiority of genomic prediction relative to pedigree was found for TSH and TPH. Some authors reported that the impact of markers in breeding decisions depends on the data structure and on how informative the pedigree and markers are (de los Campos et al. 2009; Crossa et al. 2010). As stated here, in a complex and large polyploid genome such as sugarcane, higher marker coverage should be evaluated to test this hypothesis. Moreover, the size of the training data set and small family size might be the cause of the limited accuracy obtained (514 individuals came from 98 families, with an average of 5.2 individuals per family).

Conclusions

Here we assessed the feasibility of genomic prediction methods in a commercial breeding population genotyped with SNP markers. Satisfactory accuracies were obtained by using genomic information, especially for pol percentage in sugarcane and fiber percentage in bagasse. Although phenotypic selection exhibited higher accuracies than genomic selection, the results achieved show that molecular markers could be used for prediction of genetic merit of clones, shortening the time of current selection, which requires large field experiments. Thus, the use of genomic information could be more efficient per unit of time for improvement of desirable agronomic traits in a complex polyploid crop.

Appendix

```
library(BGLR)
#Loading matrices
Y<-as.matrix(read.table("Y_514.txt",h=T)) #phenotypes
X<-as.matrix(read.table("X37024.txt",h=T)) # MM matrix
```

```
X1<-scale(X,center=TRUE,scale=TRUE)
G<-tcrossprod(X1)/ncol(X1)  # G matrix
A<-as.matrix(read.table("matrizA514.txt",h=T)) #pedigree

#Defining trait
y<-Y[,2]

#Methods
ETA<-list(MRK=list(X=X,model="BRR")) # Bayesian Ridge
Regression
ETA$MRK$model<-"BayesA" # Bayes A
ETA$MRK$model<-"BayesB" # Bayes B
ETA<-list(list(K=A,model='RKHS'))  #Pedigree-BLUP
ETA<-list(list(K=G,model='RKHS'))  #Genomic BLUP
ETA<-list(list(X=X,model='BL'))  #Bayesian LASSO
#Fitting a Pedigree + Markers model
y<-scale(y,center=TRUE,scale=TRUE)
EVD <-eigen(G)  # Computing the eigen-value decomposition
of G
ETA<-list(list(K=A2, model='RKHS'),
          list(V=EVD$vectors,d=EVD$values, model='RKHS'))
EVDG<-eigen(G); EVDA<-eigen(A)
H0<-list(PED=list(V=EVDA$vectors,d=EVDA$values,
model="RKHS")) #Setting the linear predictor
HA<-H0
HA$G<-list(V=EVDG$vectors,d=EVDG$values, model="RKHS")
COR<-matrix(nrow=nRep,ncol=2,NA)

#TRN-TST partitions
for(i in 1:nRep){
  tst<-sample(1:n,size=nTST,replace=FALSE)
  yNA<-y; yNA[tst]<-NA
  fm<-BGLR(y=yNA,ETA=H0, nIter=nIter, burnIn=burnIn)
  COR[i,1]<-cor(y[tst],fm$yHat[tst]); rm(fm)
  fm<-BGLR(y=yNA,ETA=HA, nIter=nIter, burnIn=burnIn)

  COR[i,2]<-cor(y[tst],fm$yHat[tst]); rm(fm)
}
colMeans(COR) # Comparing models using a paired t-test
mean(COR[,2]-COR[,1])
t.test(x=COR[,2],y=COR[,1],paired=TRUE,var.equal=FALSE)

#10-fold cross validation used in the models
set.seed(123) #Set seed for the random number generator
folds<-10
sets<-rep(1:10,52)[-(515:520)]
sets<-sets[order(runif(nrow(X)))]
```

```
COR.CV<-rep(NA,times=(folds+1))
DIC<-rep(NA,times=(folds+1))
names(COR.CV)<-c(paste('fold=',1:folds,sep=''),'Pooled')
w<-rep(1/nrow(X),folds) ## weights for pooled correlations
and MSE
yHatCV<-numeric()
for(fold in 1:folds)
{
  yNa<-y
  whichNa<-which(sets==fold)
  yNa[whichNa]<-NA
  prefix<-paste('PM_BL','_fold_',fold,'_',sep='')
  fm<-BGLR(y=yNa,ETA=ETA,nIter=50000, burnIn=5000,thin=10,
saveAt='BRR_FB_')
  yHatCV[whichNa]<-fm$yHat[fm$whichNa]
  w[fold]<-w[fold]*length(fm$whichNa)
  COR.CV[fold]<-cor(fm$yHat[fm$whichNa],y[whichNa])
}
COR.CV[11]<-mean(COR.CV[1:10])
COR.CV

##### Extracting results
#Godness of fit and related statistics
stats<-fm$fit
# Residual variance = % of the phenotypic variance NOT
explained by the model
var_res<-rbind(fm$varE,fm$SD.varE) # compare to var(y)

#Variance components associated with the genomic and
pedigree
# matrices
var_ped<-rbind(fm$ETA[[1]]$varU,fm$ETA[[1]]$SD.varU)
var_mark<-rbind(fm$ETA[[2]]$varU,fm$ETA[[2]]$SD.varU)

# Residual variance
# 1-VarE = % of the phenotypic variance explained by the
model
varE<-scan('PGBLUP_var4varE.dat')
#varA = estimates of the variance components associated to
the pedigree
varA<-scan('PGBLUP_var4ETA_1_varU.dat')
#varU = estimates of the variance components associated to
the markers
varU<-scan('PGBLUP_var4ETA_2_varU.dat')
varG<-varU+varA
h2<-varG/(varE+varG) #% of the phenotypic variance can be
explained by genetic factors
```

```
heritab<-rbind(mean(h2),sd(h2))
mrk_explain<-rbind(mean(varU/varG),sd(varU/varG)) #% of the
total genetic variance explained by markers
ped_explain<-rbind(mean(varA/varG),sd(varA/varG)) #% of the
total genetic variance explained by pedigree

#END
```

References

Atkin FC, Dieters MJ, Stringer JK (2009) Impact of depth of pedigree and inclusion of historical data on the estimation of additive variance and breeding values in a sugarcane breeding program. Theor Appl Genet 119:555–65. doi: 10.1007/s00122-009-1065-7

Baffa DCF, de A. Costa PM, da Silveira G, et al (2014) Path Analysis for Selection of Saccharification-Efficient Sugarcane Genotypes through Agronomic Traits. Agron J 106:1643. doi: 10.2134/agronj13.0576

Barbosa MHP, da Silveira LCI (2012) Breeding and cultivar recommendations. In: Santos F, Borém A, Caldas C (eds) Sugarcane. Bioenergy, sugar ethanol Technol. Prospect. MAPA/ACS:UFV/DEA, Brasília, DF, pp 113–119

Barbosa MHP, Resende MDV de, Dias LA dos S, et al (2012) Genetic improvement of sugar cane for bioenergy: the Brazilian experience in network research with RIDESA. Crop Breed Appl Biotechnol 87–98.

Bastos IT, Barbosa MHP, Cruz CD, et al (2003) Análise dialélica em clones de cana-de-açúcar. Bragantia 62:199–206. doi: 10.1590/S0006-87052003000200004

Bernardo R (2010) Breeding for Quantitative Traits in Plants. 2nd ed. Stemma Press, Woodbury, MN

Casella G, Berger RL (2001) Statistical inference. 2nd ed. Duxbury Press, Pacific Grove, CA

Crossa J, Beyene Y, Kassa S, et al (2013a) Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 (Bethesda) 3:1903–26. doi: 10.1534/g3.113.008227

Crossa J, Campos G de L, Pérez P, et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–24. doi: 10.1534/genetics.110.118521

Crossa J, Pérez P, Hickey J, et al (2013b) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb) 112:48–60. doi: 10.1038/hdy.2013.16

D'Hont A, Souza G, Menossi M, et al (2008) Sugarcane: A Major Source of Sweetness, Alcohol, and Bio-energy. In: Moore P, Ming R (eds) Genomics Trop. Crop Plants. Springer New York, pp 483–513

Da Silveira LCI, Brasileiro BP, Kist V, et al (2015) Selection strategy in families of energy cane based on biomass production and quality traits. Euphytica. doi: 10.1007/s10681-015-1364-9

Daetwyler HD, Calus MPL, Pong-Wong R, et al (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–65. doi: 10.1534/genetics.112.147983

Dal-Bianco M, Carneiro MS, Hotta CT, et al (2012) Sugarcane improvement: How far can we go? Curr Opin Biotechnol 23:265–270.

De Los Campos G, Hickey JM, Pong-Wong R, et al (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–45. doi: 10.1534/genetics.112.143313

De los Campos G, Naya H, Gianola D, et al (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–85. doi: 10.1534/genetics.109.101501

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics, 4th edn. Longmans Green, Harlow, Essex, UK

Federer WT (1961) Augmented designs with one-way elimination of heterogeneity. Biometrics 17:447–473.

Gouy M, Rousselle Y, Bastianelli D, et al (2013) Experimental assessment of the accuracy of genomic selection in sugarcane. Theor Appl Genet 126:2575–86. doi: 10.1007/s00122-013-2156-z

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447.

Henry RJ, Kole C (2010) Genetics, genomics and breeding of sugarcane. Science Publishers, Inc

Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci 52:146. doi: 10.2135/cropsci2011.06.0297

Hogarth DM, Wu KK, Heinz DJ (1981) Estimating Genetic Variance in Sugarcane Using a Factorial Cross Design1. Crop Sci 21:21. doi: 10.2135/cropsci1981.0011183X002100010006x

Hotta CT, Lembke CG, Domingues DS, et al (2010) The Biotechnology Roadmap for Sugarcane Improvement. Trop Plant Biol 3:75–87. doi: 10.1007/s12042-010-9050-5

Knueppel S, Rohde K (2015) Package HapEstXXR - Multi-locus stepwise regression (MSR).

Laurie CC, Doheny KF, Mirel DB, et al (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 34:591–602. doi: 10.1002/gepi.20516

Legendre B. (1992) The core/press method for predicting the sugar yield from cane for use in cane payment. Sugar J 54:2–7.

Lin Z, Hayes BJ, Daetwyler HD (2014) Genomic selection in crops, trees and forages: a review. Crop Pasture Sci 65:1177. doi: 10.1071/CP13363

Loureiro ME, Barbosa MHP, Lopes FJF, Silvério FO (2011) Sugarcane Breeding and Selection for more Efficient Biomass Conversion in Cellulosic Ethanol. In: Buckeridge MS, Goldman GH (eds) Routes to Cellul. Ethanol. Springer New York, New York, NY, pp 199–239

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Sunderland, MA

Matsuoka S, Kennedy AJ, dos Santos EGD, et al (2014) Energy Cane: Its Concept, Development, Characteristics, and Prospects. Adv Bot. doi: http://dx.doi.org/10.1155/2014/597275

Meuwissen T, Hayes B, Goddard M (2013) Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci 1:221–37. doi: 10.1146/annurev-animal-031412-103705

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Neves LG, Davis JM, Barbazuk WB, Kirst M (2013) Whole-exome targeted sequencing of the uncharacterized pine genome. Plant J 75:146–156.

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483–95. doi: 10.1534/genetics.114.164442

Peternelli LA, Ferreira FM, Rocha RB, et al (2009) Análise dos coeficientes de endogamia e de parentesco para qualquer nível de ploidia usando o pacote estatístico R. Bragantia 68:849–855. doi: 10.1590/S0006-87052009000400004

Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228. doi: 10.1007/s10681-007-9449-8

Resende MDV de (2007) SELEGEN-REML BLUP Sistema estatístico e seleção genética computadorizada via modelos lineares mistos. Embrapa Florestas, Colombo

Resende MDV de (2002) Biometric Genetics and Statistics in the Breeding of Perennial Crops (Portuguese). Brasilia: Embrapa Informação Tecnológica. Embrapa Florestas, Colombo, PR

Resende MFR, Munoz P, Resende MD V., et al (2012) Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.). Genetics 190:1503–1510. doi: 10.1534/genetics.111.137026

Schneider F (ed) (1979) Sugar analysis: ICUMSA methods. International Commission for Uniform Methods of Sugar Analysis, Peterborough, UK

Souza GM, Berges H, Bocs S, et al (2011) The Sugarcane Genome Challenge: Strategies for Sequencing a Highly Complex Genome. Trop Plant Biol 4:145–156. doi: 10.1007/s12042-011-9079-0

Stringer JK, Cox MC, Atkin FC, et al (2011) Family Selection Improves the Efficiency and Effectiveness of Selecting Original Seedlings and Parents. Sugar Tech 13:36–41. doi: 10.1007/s12355-011-0073-5

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–23. doi: 10.3168/jds.2007-0980

Viana JMS, Faria VR, Silva FF, Resende MDV de (2011) Best Linear Unbiased Prediction and Family Selection in Crop Species. Crop Sci 51:2371. doi: 10.2135/cropsci2011.03.0153

Viana JMS, Lima RO de, Faria VR, et al (2012) Relevance of Pedigree, Historical Data, Dominance, and Data Unbalance for Selection Efficiency. Agron J 104:722. doi: 10.2134/agronj2011.0358

Würschum T, Reif JC, Kraft T, et al (2013) Genomic selection in sugar beet breeding populations. BMC Genet 14:85. doi: 10.1186/1471-2156-14-85

Zeni Neto H, Daros E, Bespalhok Filho JC, et al (2013) Selection of families and parents of sugarcane (Saccharum spp.) through mixed models by joint analysis of two harvests. Euphytica 193:391–408. doi: 10.1007/s10681-013-0947-6