Theor Appl Genet (2013) 126:2575–2586
DOI 10.1007/s00122-013-2156-z

ORIGINAL PAPER

Experimental assessment of the accuracy of genomic selection in sugarcane

M. Gouy · Y. Rousselle · D. Bastianelli · P. Lecomte · L. Bonnal · D. Roques · J.-C. Efile · S. Rocher · J. Daugrois · L. Toubi · S. Nabeneza · C. Hervouet · H. Telismart · M. Denis · A. Thong-Chane · J. C. Glaszmann · J.-Y Hoarau · S. Nibouche · L. Costet

Received: 18 February 2013 / Accepted: 12 July 2013 / Published online: 2 August 2013

Abstract Sugarcane cultivars are interspecific hybrids with an aneuploid, highly heterozygous polyploid genome. The complexity of the sugarcane genome is the main obstacle to the use of marker-assisted selection in sugarcane breeding. Given the promising results of recent studies of plant genomic selection, we explored the feasibility of genomic selection in this complex polyploid crop. Genetic values were predicted in two independent panels, each composed of 167 accessions representing sugarcane genetic diversity worldwide. Accessions were genotyped with 1,499 DArT markers. One panel was phenotyped in Reunion Island and the other in Guadeloupe. Ten traits concerning sugar and bagasse contents, digestibility and composition of the bagasse, plant morphology, and disease resistance were used. We used four statistical predictive models: bayesian LASSO, ridge regression, reproducing kernel Hilbert space, and partial least square regression. The accuracy of the predictions was assessed through the correlation between observed and predicted genetic values by cross validation within each panel and between the two panels. We observed equivalent accuracy among the four predictive models for a given trait, and marked differences were observed among traits. Depending on the trait concerned, within-panel cross validation yielded median correlations ranging from 0.29 to 0.62 in the Reunion Island panel and from 0.11 to 0.5 in the Guadeloupe panel. Cross validation between panels yielded correlations ranging from 0.13 for smut resistance to 0.55 for brix. This level of correlations is promising for future implementations. Our results provide the first validation of genomic selection in sugarcane.

Communicated by J. Crossa.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-013-2156-z) contains supplementary material, which is available to authorized users.

M. Gouy · A. Thong-Chane
eRcane, 97494 Sainte-Clotilde, La Réunion, France

M. Gouy · Y. Rousselle · H. Telismart · S. Nibouche ·
L. Costet (✉)
Cirad, UMR PVBMT, 97410 Saint-Pierre, La Réunion, France
e-mail: laurent.costet@cirad.fr

M. Gouy
Université de La Réunion, UMR PVBMT, 97410 Saint-Pierre,
La Réunion, France

D. Bastianelli · P. Lecomte · L. Bonnal
Cirad, UMR SELMET, 34398 Montpellier, France

D. Roques · J.-C. Efile · S. Rocher · L. Toubi · J.-Y. Hoarau
Cirad, UMR AGAP, 97170 Petit Bourg, Guadeloupe, France

S. Rocher
Université des Antilles et de la Guyane,
97157 Pointe-à-Pitre, Guadeloupe, France

J. Daugrois
Cirad, UMR BGPI, 97170 Petit Bourg,
Guadeloupe, France

S. Nabeneza
Cirad, UMR SELMET, 97410 Saint-Pierre,
La Réunion, France

C. Hervouet · M. Denis · J. C. Glaszmann
Cirad, UMR AGAP, 34398 Montpellier, France

Introduction

Sugarcane (*Saccharum* spp.) is a clonally propagated industrial crop cultivated for its high sucrose content as well as for bioenergy purposes. Sugarcane supplies about 80 % of the world's sucrose and is cultivated in almost 100 tropical or sub-tropical countries (FAOSTAT 2012). Modern cultivars are derived from interspecific crosses between *S. officinarum* and *S. spontaneum*. They have a highly polyploid genome, highly heterozygous and frequently aneuploid. Modern sugarcane cultivars have between $2n = 100$ and $2n = 130$ chromosomes distributed in about 12 homologous groups for a total size of 10 Gb (Grivet and Arruda 2001; D'Hont 2005). Chromosome behavior at meiosis displays mainly bivalents, but pairing affinity within a given homology class shows complex patterns due to subsets of chromosomes exhibiting variable ranges of preferential mutual pairing (Jannoo et al. 2004) and possible instances of systematic disomic pairs (Hoarau et al. 2001). This complex sugarcane genome organization makes trait inheritance difficult to analyze. Currently, sugarcane breeding programs still rely on phenotypic selection in the framework of quantitative genetic approaches, usually based on combined mass and family selection strategies. Improvement of agronomic traits related to yield and disease resistance require large experiments that last for several crop cycles (Cheavegatti-Gianotto et al. 2011). About 7–10 years of field experiments are needed to identify elites for further multi-location pre-commercial tests (Del Blanco et al. 2010; Scortecci et al. 2012). Marker-assisted selection (MAS) approaches could have been valuable strategies for sugarcane breeders like in other crop models (Collard and Mackill 2008; Xu and Crouch 2008). However, the complexity of the sugarcane genome makes marker-trait association studies challenging. In the past two decades, numerous quantitative trait loci (QTLs) studies relative to yield components have been published on sugarcane based on biparental progenies (Hoarau et al. 2002; Ming et al. 2002; Reffay et al. 2005; Aitken et al. 2006, 2008; Piperidis et al. 2008; Pinto et al. 2010). More recently, genome-wide association studies (GWAS) were developed to identify QTLs relative to several agronomic traits (Wei et al. 2006, 2010). A lot is expected from this latest strategy, which is facilitated by the persistence of linkage disequilibrium (LD) observed within the first 5 cM among modern cultivars due to recent bottlenecks in breeding schemes (Jannoo et al. 1999; Raboin et al. 2008). However, all these QTL studies revealed common features: a modest phenotypic effect of most QTLs and frequent lack of repeatability across environments or crop cycles. In such a context, validating common QTLs between studies is not easy for agronomic traits related to yield components (Piperidis et al. 2008). Up to now, the use of molecular markers in sugarcane breeding programs has only been conceivable to tag a very few major resistance genes related to diseases (Aljanabi et al. 2007; Costet et al. 2012a, b; Glynn et al. 2013).

Genomic selection, an approach introduced by Meuwissen et al. (2001), is a novel method for selecting individuals in breeding programs, and is suitable for the improvement of complex traits requiring long field experiments (Heffner et al. 2009; Lorenz et al. 2011). Its principle is to predict the phenotypic performance of individuals, in terms of breeding value or total genetic value, on the basis of their genome-wide genotypic data, by using a predictive model previously calibrated with a representative phenotyped and genotyped 'training population'. Genomic selection exploits the whole marker information by simultaneously estimating the effect of each marker across the entire genome to predict the genetic value of individuals (Meuwissen et al. 2001). Unlike conventional marker-assisted selection, genomic selection does not rely on a subset of significant markers. Therefore, genomic selection models should have the ability to capture more of the genetic variation by taking QTLs with small-effect into account. Genomic selection was first applied for livestock breeding. Added value brought by genomic selection predictions, demonstrated through theoretical simulations (Meuwissen et al. 2001; Calus et al. 2008; Solberg et al. 2008) or empirical evidence (Luan et al. 2009; Moser et al. 2009; VanRaden et al. 2009) now opens a new area of development in animal breeding programs. In plant breeding, numerous simulated and experimental genomic selection studies have been recently published, on cereal crops such as wheat, barley, or maize (Lorenzana and Bernardo 2009; Zhong et al. 2009; Crossa et al. 2010; Lorenz et al. 2011; Heslot et al. 2012; Guo et al. 2012; Rutkoski et al. 2012) and on forest trees such as loblolly pine (Resende et al. 2012b) and eucalyptus (Grattapaglia et al. 2011; Resende et al. 2012a; Denis and Bouvet 2013). Given the increasing access to high-throughput genotyping at more affordable costs, opportunity to test genomic selection approaches should expand on a larger number of plant species. To date, this approach has never been tested in sugarcane.

The objective of our study was to evaluate genomic selection in sugarcane using four prediction models and ten quantitative traits of agronomic interest. Predictions were based on the genetic value of two independent panels composed of accessions representing the diversity of germplasm used in breeding programs worldwide. The accuracy of the genomic selection models was assessed by cross validations within panels and between panels.

Materials and methods

Plant material

Two independent sugarcane panels were used in this study, REU (phenotyped in Reunion Island) and GUA (phenotyped in Guadeloupe), each composed of 167 accessions. The two panels are subsamples of the REUb and GUA panels, described in detail in Costet et al. (2012a), chosen to be totally exclusive from one another (no common materials). The 334 accessions were modern cultivars and breeding material derived more or less recently from over 30 breeding centers around the world. The REU panel was maintained in Reunion Island at the eRcane and CIRAD experimental stations. The GUA panel was maintained in Guadeloupe at CIRAD Roujol experimental station.

Field trials

The REU and GUA panels were phenotyped using a total of seven field trials covering several crop cycles. In Reunion Island, the REU panel was phenotyped in four locations: Bassin-Martin, La Mare, Vue-Belle, and Le Gol. In Guadeloupe, the GUA panel was phenotyped in three locations: Roujol-1, Roujol-2, and Godet. The La Mare trial was planted during the austral summer (October 2010), while the Bassin-Martin (April 2010), Vue-Belle (April 2010), and Le Gol (June 2006) trials were planted before austral winter. All three locations in Guadeloupe were planted before winter (September) in different years (Roujol-1: 2005; Roujol-2: 2007; Godet: 2010). Standard cultivation practices (fertilization, weeding) were used. All trials were irrigated to avoid water stress. The Reunion Island trials used an alpha-lattice design with three replications and 20 blocks in each replication. In Guadeloupe, we used complete randomized block designs with three replications.

Phenotypic data

Details on the trials and crop cycles used to phenotype each trait are given in online resource 1. The two panels were phenotyped for a total of ten agronomic traits related to cane yield (morphological and technological traits): lignocellulose composition of the bagasse, and disease resistances. Morphological traits measured were stalk diameter (SD) and number of millable stalks (SN) both measured at harvest. However, stalk height could not be measured because the majority of our accessions flowered at harvest which prevented this trait measurement without bias. Stalk diameter was the mean of three measures (taken at the bottom, mid-height, and top of the stalk) made on nine randomly chosen stalks per plot. The number of stalks was counted in one row per elementary plot and expressed as the number of stalks per square meter. Technological traits were juice brix (BR) and bagasse content (BC). Brix is a measure of soluble solids in the sugarcane juice, expressed as a percentage of solids by weight (% w/w). In Guadeloupe, BR was measured in the field with a handheld refractometer on the juice of a sampling punch taken at half-height of seven randomly chosen stalks per elementary plot. In Reunion Island, nine randomly chosen stalks were crushed in eRcane laboratories. A 500-g sample of pulp was then pressed using the hydraulic press method (Hoarau 1969). BR was measured with a refractometer on the collected juice. In both islands, bagasse content was estimated using the same hydraulic press method (Hoarau 1969) on the basis of the ratio of the fresh weight of the cake (after juice extraction) to the fresh weight of the pulp. The sequential method of Van Soest et al. (1991) was used to describe the fiber fractions. This method provides an estimate of total fiber (NDF, neutral detergent fiber), lignocelluloses (ADF, acid detergent fiber), and lignin content (ADL, acid detergent lignin). In our study, we used the variables ADF and ADL expressed as a proportion of total fiber (NDF), to characterize the quality of the fiber rather than the absolute values of these compounds which rely on sucrose content (Barrière et al. 2010). In addition, in vitro NDF digestibility (IVNDFD) was determined using an enzymatic method with pepsin and cellulase (Aufrère et al. 2007). IVNDFD represents the potential degradation of fiber in ruminants and is, therefore, used as an index of biomass quality.

ADF, ADL, and IVNDFD were determined using near-infrared spectroscopy (NIRS) after calibration with reference analyses, as described in Sabatier et al. (2012). The data for both REU and GUA panels were obtained using the same analytical procedures and NIRS equations.

Disease resistance phenotyping focused on three of the main sugarcane diseases in the world: the fungal smut disease, caused by *Sporisium scitaminea,* brown rust, caused by *Puccinia melanocephala*, and yellow leaf disease caused by sugarcane yellow leaf virus (SCYLV, Polerovirus). Smut resistance (SM) was assessed in trials on which plants were artificially inoculated, whereas rust resistance (RST) and yellow leaf disease resistance (SCYLV) were assessed under natural infection. Scoring of the rust reaction was performed according to the method of Costet et al. (2012b). Based on these results, only accessions that do not carry the major resistance gene *Bru*1 that confers complete resistance were used in the analysis: 111 accessions from the GUA panel and 85 from the REU panel. This selection was made to allow prediction of quantitative resistance, independently of resistance conferred by *Bru*1. SCYLV was detected by tissue blot immunoassay (Schenck et al. 1997) on the first fully emerged leaf of ten stalks randomly sampled in each elementary plot and used to compute the incidence of SCYLV. To evaluate SM, accessions were inoculated at planting by dipping the cuttings in a suspension of 5×10^6 spores ml^{-1} for 20 min. Spores were isolated from whips collected in fields in Reunion Island or in Guadeloupe. Smut incidence was measured by the cumulated number of whips per elementary plot during three crop cycles.

Genotypic data

Accessions were genotyped using DArT markers (Heller-Uszynska et al. 2011). DNA of the 334 genotypes was sent for genotyping to the private company Diversity Arrays Technology Pty Ltd. A total of 1,758 markers were common to the two panels. Low-, or high-frequency markers (<0.05 and >0.95) or with more than 10 % missing data were removed. Each marker's within panel marker frequency was used to impute missing data (Heslot et al. 2012). The number of DArT markers after edition was 1,499.

Statistical methods

Analyses of quantitative traits

The phenotypic data of each panel were analyzed separately using mixed linear models and generalized mixed linear models to estimate variance components and genetic values (GV). The general mixed linear model was used for morphological (SD, SN), technological (BR, BC), and lignocellulose traits (ADL, ADF, IVNDFD) traits and can be written as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z_1 b} + \mathbf{Z_2 c} + \mathbf{Z_3 cl} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of phenotypic observations for each trait, $\boldsymbol{\beta}$ a vector of fixed effects related to the experimental design including the fixed effects of location, crop cycle and replication, $\mathbf{b}$ the vector of random incomplete block effects within each replication $\sim N\left(0, \mathbf{I}\sigma_{\mathrm{b}}^2\right)$, $\mathbf{c}$ the vector of random effects of clones $\sim N\left(0, \mathbf{I}\sigma_{\mathrm{c}}^2\right)$, $\mathbf{cl}$ the vector of random effects of interaction between genotypes and location or crop cycle $\sim N\left(0, \mathbf{I}\sigma_{\mathrm{cl}}^2\right)$, and $\mathbf{e}$ the vector of residual error of the model $\sim N\left(0, \mathbf{I}\sigma_{\mathrm{e}}^2\right)$. $\mathbf{X}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$, and $\mathbf{Z_3}$ are incidence matrix, and $\mathbf{I}$ the identity matrix. Linear mixed models were fitted using the lme4 package (Bates et al. 2011) and convergence was checked for each analysis. Because of their non-Gaussian distributions, the three disease-related traits (RST, SCYLV and SM) could not be analyzed using linear models. We, therefore, used the MCMCglmm package for R (Hadfield 2010) that implements Markov Chain Monte Carlo routines to fit multi-response generalized linear mixed models. For rust resistance, which is an ordinal response, we used the standard threshold model (Sorensen and Gianola 2002). For SCYLV resistance, we used a binomial response with a logit link function, and for smut resistance an over-dispersed Poisson response with a log link function. An inverse Wishart prior was used for the variance components. This prior distribution takes two scalar parameters V and ν. Because we do not have any prior knowledge, a non-informative prior was used, for both genetic and residual variances, by putting $V = 1 \times 10^{-16}$ and $\nu = -2$ (Hadfield 2012). Each model was run for 50,000 Markov chain Monte Carlo (MCMC) simulation iterations. We discarded the first 15,000 cycles as burn-in after checking the stability of posterior values. We checked for convergence of model parameter estimates by inspecting trace plots of the MCMC iterations and autocorrelations plots. We chose a thinning interval of 10 iterations, which resulted in 3,500 posterior distribution samples of model parameter estimates. Most of genomic prediction models used in this study rely in the normality assumption. Histograms of estimated genetic values were drawn in order to assess normality. Broad sense heritability at the experimental design level and coefficients of genetic variation were calculated for the normally distributed traits according to Gallais (1990). Because the variance components of the three diseases traits were transformed in the link function scale (log for SM, logit for SCYLV or probit for RST), heritabilities of these traits were not estimable. We extracted the random vector of genotypic effects from the models for each accession panel and considered these data as the genetic values for the rest of the study.

Linkage disequilibrium and genetic diversity in the panels

The polyploidy of sugarcane associated with the dominant nature of the markers used prevented the computation of allele frequency on which the classical measures of LD (D', r^2, d^2) are based. Following Raboin et al. (2008), we, therefore, used pairwise Fisher exact tests to detect associations among markers. The probabilities of these tests were plotted between the two panels to compare the association pattern of markers. The genetic diversity of the panels was compared by principal component analysis (PCA) performed on the whole 334 accessions using the 1,499 markers previously standardized.

Models used for genomic selection

Four predictive models were compared: two parametric models, ridge regression (RR) (Hoerl and Kennard 1970) and bayesian LASSO (BL) (Park and Casella 2008); one semi-parametric model, reproducing kernel Hilbert spaces (RKHS) (Schölkopf and Smola 2002); and a non-parametric model, partial least square regression (PLSR) (Wold

2001). These models were fitted using the R software (R Core Team 2013).

The penalized method RR and the Bayesian method BL are two shrinkage methods commonly used for genomic selection. They differ by the extent and the kind of shrinkage: in RR, the shrinkage is homogenous across markers while in BL it is heterogeneous. A clear description of these methods and their equivalences is given in de los Campos et al. (2013). We assessed BL using the BLR package version 1.3 (de los Campos and Perez 2010). The BL method shrinks more estimates of marker effects that are close to zero and less those with high effects. The marginal prior distribution of marker effects is a double exponential. Estimates of marker effects ($\hat{\beta}$) are obtained by solving the constrained optimization problem below:

$$\hat{\beta} = \min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - x_i'\boldsymbol{\beta}\right)^2 + \lambda(t) \sum_{j=1}^{p} \left|\beta_j\right| \right\}$$

where y_i is the observed value of individual i, x_i is the ith row of the markers matrix, $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients, λ the regularization parameter controlling the trade-offs between goodness of fit, t is an arbitrary positive constant, and β_j the estimate effect of the jth marker. Hyper-parameters were chosen based on the guidelines of Pérez et al. (2010). The regularization parameter λ and the scale parameter of the residual variance (S_e) were computed as follows:

$$\hat{\lambda} = \sqrt{2V_e V^{-1} \sum_{j}^{p} \bar{x}_j^2}$$

$$S_e = V_e(df_e + 2),$$

where $\bar{x}_j$ denotes the average value of the jth column of $\mathbf{X}$ the genotyping matrix. V_e and V correspond to the residual and genetic variance, respectively, obtained using mixed models described in a previous section (Table 1). Degree of freedom was chosen at $df_e = 4$ (Resende et al. 2012b) to guarantee finite variances (Pérez et al. 2010).

The Gibbs sampler was run for 30,000 cycles and the first 5,000 cycles were discarded as burn-in. Posterior of genetic and residual variances were checked and autocorrelations plots were drawn to ensure that models converged. The thin parameter was chosen at 10.

The package rrBLUP version 3.8 (Endelman 2011) was used to perform the RR and RKHS methods.

RR was first proposed for marker-assisted selection by Whittaker et al. (2000). This penalized method performs an extent of shrinkage that is homogeneous. Estimates of marker effects ($\hat{\beta}$) are obtained as follows:

Table 1 Summary statistics from linear mixed models of phenotypic data used for genomic selection

Traits	Panel	$\hat{\sigma}_G^{2}$ [a]	$\hat{\sigma}_e^{2}$ [b]	H² [c]	CV_g [d]	Mean ± SEM
Morphological traits						
SN (stalk/m²)	REU	79.16	99.13	0.80	23	39.21 ± 0.40
	GUA	118.48	51.14	0.90	26.1	41.66 ± 0.42
SD (mm)	REU	6.29	3.52	0.89	9.8	25.93 ± 0.10
	GUA	7.85	0.54	0.96	10.2	26.86 ± 0.07
Technological traits						
BR (%)	REU	0.98	1.29	0.83	5.4	17.94 ± 0.05
	GUA	1.63	0.58	0.88	6.2	20.49 ± 0.06
BC (%)	REU	2.04	1.59	0.89	8.3	17.52 ± 0.05
	GUA	1.81	0.88	0.71	8.1	15.98 ± 0.07
Lignocellulose traits						
ADL (%)	REU	0.32	0.37	0.84	4.9	11.77 ± 0.03
	GUA	0.68	0.092	0.87	6.9	11.88 ± 0.04
ADF (%)	REU	1.2	1.9	0.78	1.7	62.62 ± 0.06
	GUA	2.1	0.22	0.86	2.3	62.01 ± 0.07
IVNDFD (%)	REU	2.84	4.97	0.79	19.4	8.70 ± 0.13
	GUA	6.99	0.84	0.87	27.8	9.52 ± 0.15
Disease traits						
RST (score)	REU	4.12	0.51	–	–	2.95 ± 0.08
	GUA	13.17	0.48	–	–	3.19 ± 0.08
SCYLV (%)	REU	28.58	0.03	–	–	72.33 ± 1.85
	GUA	10.90	0.12	–	–	76.57 ± 1.12
SM (whip/m²)	REU	5.06	1.13	–	–	5.91 ± 0.69
	GUA	5.99	1.76	–	–	5.48 ± 0.51

– not estimated

SN stalk number, *SD* stalk diameter, *BR* brix, *BC* bagasse content, *ADL* acid detergent lignin as a percentage of neutral detergent fiber, *ADF* acid detergent fiber as a percentage of neutral detergent fiber, *IVNDFD* in vitro neutral detergent fiber digestibility of the bagasse, *RST* rust resistance, *SCYLV* yellow leaf disease resistance, *SM* smut resistance, *SEM* standard error of the mean

[a] Genetic variance

[b] Residual variance

[c] Broad sense heritability at the experimental design level

[d] Coefficient of genetic variation

$$\hat{\beta} = \left(\mathbf{X}\mathbf{X}' + \lambda\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}$$

where $\hat{\beta}$ is the vector of estimates of marker effects, **X** is the matrix allocating all genotypes to phenotypes, **I** the identity matrix, **y** is the vector of phenotypes, and λ is the ridge parameter. Choice of λ was made using a fast spectral algorithm for mixed models, included in the rrBLUP package. For these two parametric methods (BL, RR), predicted genetic values (PGV) were obtained as follows:

$$\mathbf{PGV} = \mathbf{X}\hat{\beta}$$

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

The RKHS method is supposed to capture non-additive effects (Endelman 2011). PGV were obtained using the exponential kernel $K_{i,j}$ included in the package rrBLUP. This kernel can be expressed as $\mathrm{K}_{i,j} = \exp\left(-\frac{D_{i,j}}{\theta}\right)$, where $D_{i,j}$ is the Euclidian distance between genotype i and j normalized to the interval [0, 1], and θ is a scale parameter that controls how genetic covariance decays with genetic distance. Restricted maximum likelihood was used to identify the optimal scale parameter.

For the PLSR method estimations, we used the R package pls (Mevik et al. 2011). The PLSR method maximizes the empirical covariance between predictors and response vectors by searching for linear combinations (considered as latent variables or components) of the predictors. These latent variables are orthogonal to avoid the problem of multicollinearity. The model can be written as follows:

$$\mathrm{g}(\mathbf{x}_i) = \sum_{l=1}^{k} t_{il}\beta_l$$

where $\mathrm{g}(\mathbf{x}_i)$ is the predicted genetic values (PGV) of genotype i, t_{il} is the lth latent variable of the ith genotype, and β_l is the effect associated with the lth latent variable. We used the kernel algorithm of Dayal and MacGregor (1997) to calculate latent variables. The number of latent variables k, used in the predictive model, can be determined by cross validation: the training panel is randomly split into five segments of which four serve as the training panel to predict the remaining fifth. We repeated the cross validation for $k = 1$–50, 5,000 times and computed the root mean square error of prediction (RMSEP) for each k value. Ultimately, the number of components k chosen to predict the other panel was the one that gave the lowest average RMSEP.

Accuracy of genomic selection predictions

We considered the accuracy of genomic selection prediction as the correlation between the genetic values predicted by genomic selection (PGV) and the observed genetic values (GV). We used the Pearson correlation coefficient as the measure of the prediction accuracy. The first approach used to evaluate the accuracy of genomic selection was a fivefold cross-validation within each panel. For each data set, the 167 accessions were randomly split into five subsets of which four were used as the training set to predict the remaining fifth. The random sampling of the training and validation sets was repeated 500 times. Models were compared using the median, the fifth, and the 95th percentiles of the correlation coefficient values.

The second approach used to evaluate the accuracy of genomic selection was cross-validation between panels. For each model, predictions were made by interchanging the training and validation panels: the Guadeloupe panel for training to predict the Reunion Island panel and vice versa. The four GS methods were compared with both cross-validation methods. Hyper-parameters and number of iterations were similar for each approach.

Estimates of marker effects were compared with two different genomic selection methods: BL and RR. We focused on four different traits: BR, SN, ADL as a percentage of neutral detergent lignin, and SCYLV incidence. These estimates were plotted to compare homogeneous shrinkage by ridge regression with the selective heterogeneous shrinkage made by BL.

Results

Phenotypic data

A broad range of variation was observed. For all traits, genotypic variance was significantly higher than zero within both panels ($p < 0.01$). Broad sense heritabilities (H^2) were high and varied from 0.71 to 0.96 for traits analyzed with general linear mixed models (Table 1). Coefficients of genetic variation (CV_g) revealed marked differences among traits: ADF in percent of neutral detergent fiber, ADL in percent of neutral detergent fiber, BR, and BC had the lowest values, below 10 %, while fiber digestibility estimated with IVNDFD had the highest coefficient with 27.8 %, suggesting a large variability available for this biomass quality trait. Histograms of GV exhibit a Gaussian distribution for most of the traits except for disease resistance traits (Online resource 2). For all traits, mean values and coefficients of genetic variation were similar between the two panels.

Linkage disequilibrium and genetic diversity of the panels

Probabilities of pairwise marker associations exhibited a similar pattern in the two panels (Fig. 1). This suggested

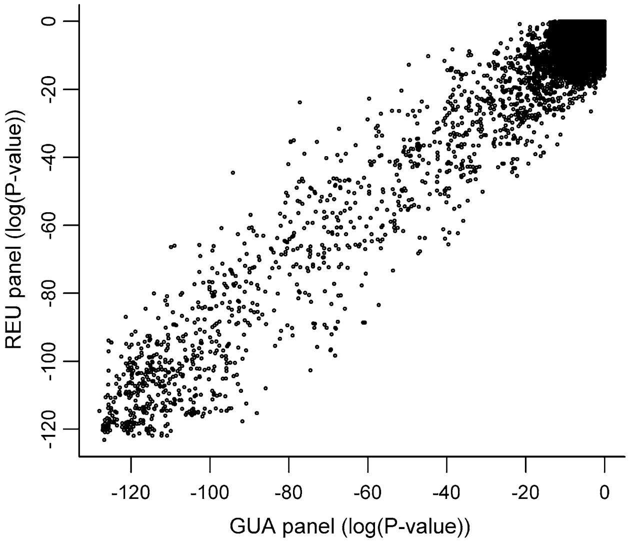

Fig. 1 Comparison $\log_{10}$ (*P* values) of pairwise Fisher exact tests between 1,499 markers within each panel

that associations between markers were similar in the two panels, indicating a similar pattern of LD in the two panels. The principal component analysis of the global sample led to major eigenvectors that bore a small proportion of the variation, in accordance with the large number of variables. Axis 1 bore 3.6 % of the total inertia, while the following axes showed a slow and continuous decrease. The first three axes of the principal component analysis summarized about 10 % of the total marker inertia (Fig. 2). The projection of the 334 accessions on the first three components revealed no disjunction of the two panels and a similar organization of the genetic diversity in the REU and in the GUA panels.

Accuracy of genomic selection predictions

In the within-panel cross validation, the four genomic selection methods used exhibited similar median correlation values between predicted and observed genetic values whatever the trait (Fig. 3, Online resource 3). Median correlations between predicted and observed genetic values ranged from 0.29 to 0.62 for the REU panel and from 0.11 to 0.50 for the GUA panel. Seven traits were better predicted in the REU panel than in the GUA panel: SN, BR, BC, ADF, RST, SCYLV, and SM. Two traits were better predicted in the GUA panel than in the REU panel: ADL and IVNDFD. The best accuracy was observed for the BR with a correlation of 0.62 between observed and predicted genetic values. The medians of the correlation values within the 5th—95th percentiles, obtained with 500 random samplings, covered a wide range of values (Online resource 3).

In the cross validation between panels, all four predictive models yielded similar correlations between predicted and observed genomic selection whatever the trait (Fig. 4, Online resource 4). Marked differences were observed between traits, with correlations ranging from 0.13 (non-significant) to 0.55 (significant). Morphological traits (SD and SN), were predicted with correlations ranging from 0.31 to 0.47. Accuracy of predictions of bagasse lignocellulose traits was generally low to moderate (ranging from 0.16 to 0.33) with best accuracies observed for the ADL in percent of neutral detergent fiber. Technological traits show low to moderate accuracies. BC was predicted with correlations ranging from 0.24 to 0.34 and the BR with correlations ranging from 0.42 to 0.55. Among the three disease resistances studied, SM was predicted with the lowest accuracy, from

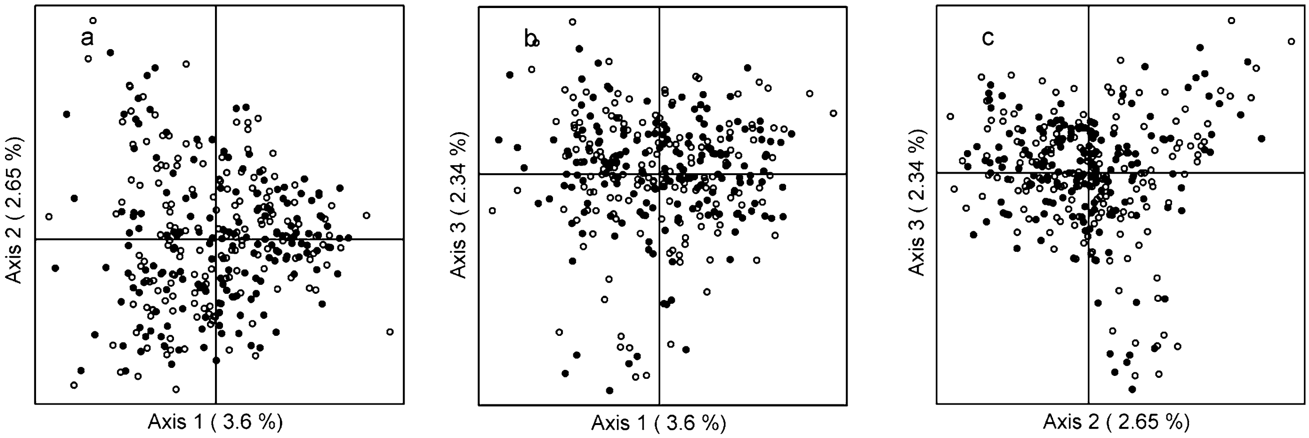

Fig. 2 Principal component analysis of 334 accessions genotyped with 1,499 independent DArT markers. **a** Biplot of the first two components. **b** Biplot of the first and the third components. **c** Biplot of the second and third components. Genotypes of the GUA (Guadeloupe) panel are identified by *circles* and genotypes from REU (Reunion Island) panel by *dots*

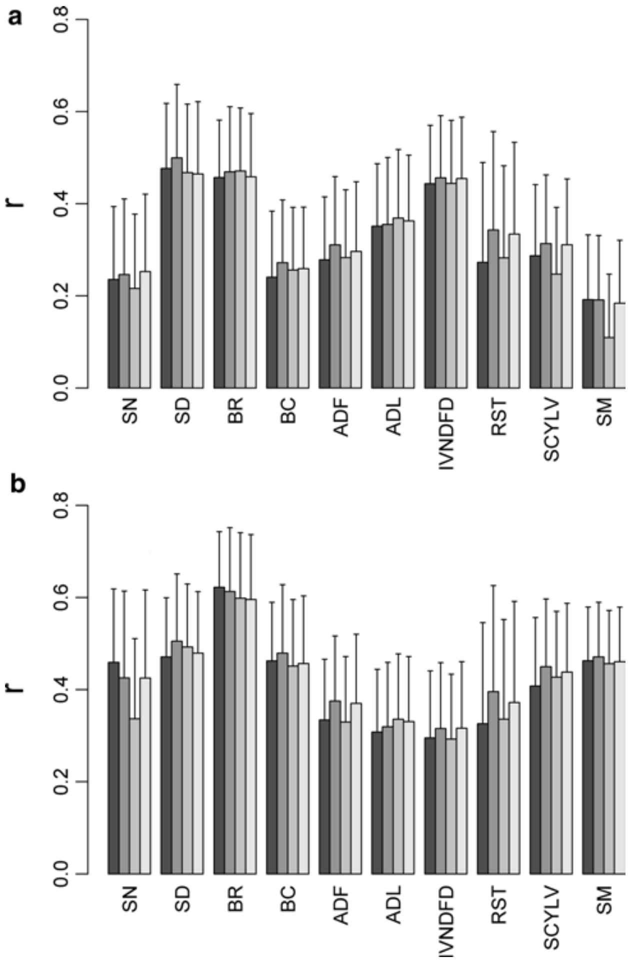

Fig. 3 Median correlations (Pearson's coefficient) between observed genetic values (GV) and predicted genetic values (PGV) in a fivefold within-panel cross validation. Four genomic selection methods were compared. From darkest to lightest: bayesian LASSO, reproducing kernel Hilbert space, partial least square regression, and ridge regression. Ten traits were predicted: stalk diameter (SD), stalk number (SN), brix of the juice (BR), bagasse content (BC), acid detergent lignin as a percentage of neutral detergent fiber (ADL), acid detergent fiber as a percentage of neutral detergent fiber (ADF), in vitro neutral detergent fiber digestibility of the bagasse (IVNDFD), rust resistance (RST), yellow leaf disease resistance (SCYLV), and smut resistance (SM). **a** Cross validation within the GUA panel. **b** Cross validation within the REU panel. For RST, we focused on accessions which do not carry the major resistance gene *Bru*1. *Vertical lines* over the bars represent absolute value of the standard deviations

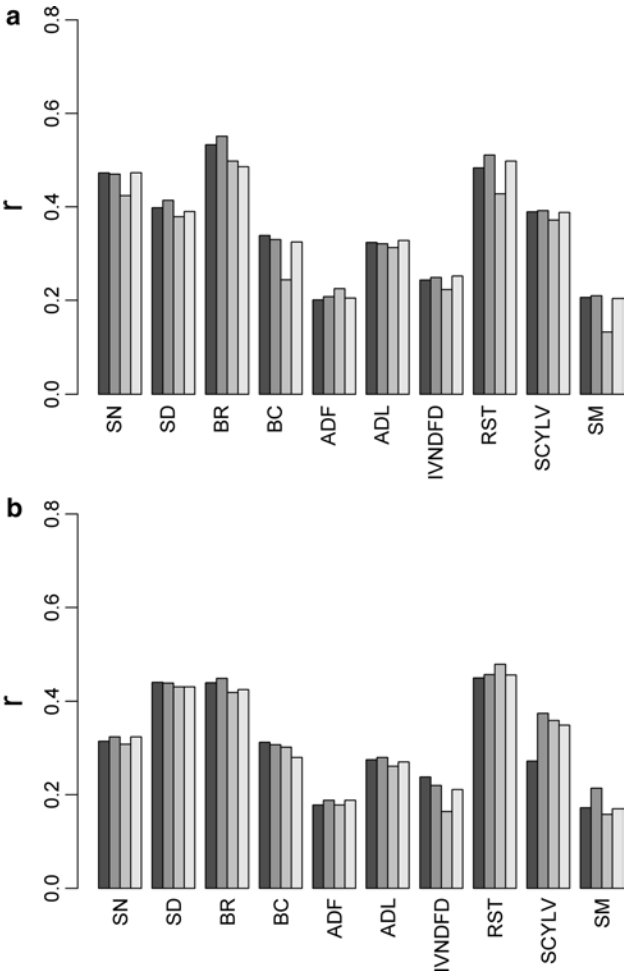

Fig. 4 Correlation coefficients (Pearson's coefficient) between observed genetic values (GV) and predicted genetic values (PGV) obtained using cross validation between two independent panels. Four genomic selection methods were compared. From darkest to lightest: bayesian LASSO, reproducing kernel Hilbert space, partial least square regression and ridge regression. Ten traits were predicted: stalk diameter (SD), stalk number (SN), brix of the juice (BR), bagasse content (BC), acid detergent lignin as a percentage of neutral detergent fiber (ADL), acid detergent fiber as a percentage of neutral detergent fiber (ADF), in vitro neutral detergent fiber digestibility of the bagasse (IVNDFD), rust resistance (RST), yellow leaf disease resistance (SCYLV), and smut resistance (SM). **a** The GUA panel was used as training population to predict the REU panel. **b** The REU panel was used as training population to predict the GUA panel. For RST, we focused on accessions that do not carry the major resistance gene *Bru*1

0.13 (non-significant) to 0.21. SCYLV was predicted with a correlation ranging from 0.27 to 0.39. For RST, the quantitative part of the resistance was predicted with a correlation ranging from 0.43 to 0.51. All correlation coefficients are detailed in Online resource 4. Estimates of marker effects obtained with BL and RR were compared (Fig. 5). Marker effects were closely correlated between the two methods. For three of the four traits studied here (BR, SN, and ADL), the BL method has marker effects more shrunk toward zero than the RR method.

Discussion

Our results show that the level of genomic characterization that we applied allows the prediction of several agronomic traits between two sugarcane populations evaluated in two

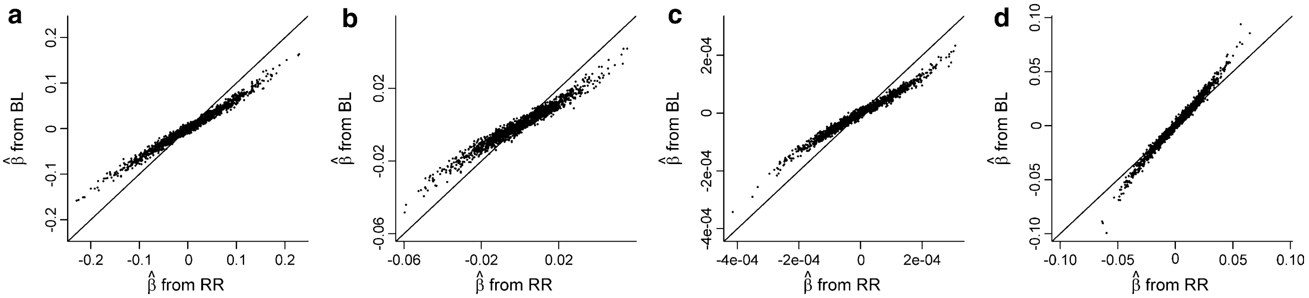

Fig. 5 Comparisons of marker effects estimated with ridge regression (RR) and bayesian LASSO (BL) for the GUA panel. Biplot of estimates of marker effects ($\hat{\beta}$) with ridge regression (RR) and bayesian LASSO, for **a** stalk number (SN), **b** brix of the juice (BR), **c** acid detergent lignin (ADL) and **d** yellow leaf disease resistance (SCYLV)

different geographic regions. This is a striking result for such a complex genome crop.

The conclusions are unlikely to be the result of a choice of statistical model. We assessed four models representative of the range of methods currently available for genomic selection, including a diverse statistical backgrounds (parametric, semi-parametric and nonparametric models, frequentist versus bayesian approaches) and validation methods (cross validation within or between panels). No major difference in prediction accuracy was observed between the methods, while we observed marked differences between traits, with accuracies ranging from 0.11 (non-significant) to 0.62 (highly significant). The congruence between models confirms the general observations of Moser et al. (2009) and Heslot et al. (2012). The study of Moser et al. (2009) compared five regression methods applied for two traits: least square regression, bayesian ridge regression, random regression best linear unbiased prediction, PLSR, and nonparametric support vector regression. They showed that the accuracy of these methods was similar except for least square regression which does not use information from all markers but from a subset of selected single-nucleotide polymorphisms (SNP). Heslot et al. (2012) tested 11 genomic selection methods on eight empirical datasets (*Arabidopsis thaliana*, wheat, barley and maize datasets). They observed that the accuracies of genomic selection models were similar whatever the trait and the population used. However, they showed that empirical Bayes and elastic net methods produced a marker effect distribution with an extremely high kurtosis, whereas the ridge regression methods produced marker effect distribution that rarely departed from a normal distribution excess kurtosis. In our study, BL has shrunk estimated marker effects more toward zero than RR, as expected (Lorenz et al. 2011). Despite these differences, both methods give similar results. When a trait is controlled by more than 20 QTLs, the advantages of heterogeneous shrinkage over homogeneous shrinkage can disappear (Zhong et al. 2009). In sugarcane, genetic studies revealed a large number of small effect QTLs for yield components (Hoarau et al. 2002; Aitken et al. 2006; Wei et al. 2010) and for quantitative disease resistance such as resistance to smut (Raboin et al. 2001, 2003). The large number of QTLs controlling the traits we analyzed may explain why bayesian methods did not outperform RR method. Moreover, as pointed out by Jannink et al. (2010) in his review, bayesian methods are able to improve the accuracy of predictions only when markers were strongly associated with QTLs. Considering the important size of the sugarcane genome, we could believe that our marker coverage do not permit strong associations between markers and QTLs. A thousand and a half markers is certainly insufficient for a genome that can over 17,000 cM (Hoarau et al. 2001). RR, RKHS, and BL rely on normality assumption whereas the partial least square does not. Histograms of genetic values for disease resistance traits (Online resource 2) deviate from normality, which, therefore, could have disadvantaged methods based on this assumption. Despite this deviation, we did not observe significant differences in prediction accuracy of PLSR and others methods.

Our levels of accuracy are of the same order of magnitude as those of other studies based on empirical data with conditions close to ours (Crossa et al. 2010; Rutkoski et al. 2012). We observed correlations ranging from 0.13 to 0.55 for cross validation between panels and from 0.11 to 0.62 for cross validation within each panel.

Diverse features may be important for the success of a trait-prediction experiment through genome-wide genotyping. Persistence of LD across panels is an important prerequisite that affects prediction accuracy (Meuwissen et al. 2001; De Roos et al. 2008). As demonstrated by simulation studies, marker density relative to population's effective size (N_e) and LD in breeding material are main factors determining accuracy levels of genomic selection prediction (Calus et al. 2008; Solberg et al. 2008). The accuracy of genomic selection increased with the average r^2 between adjacent markers. We have shown that our two sugarcane panels shared a similar genetic diversity pattern. Linkages between marker pairs displayed a similar pattern in the two panels. The recent bottleneck in the sugarcane breeding history suggests a small N_e value and may be the cause of the high level of LD (5 cM) (Jannoo et al. 1999; Raboin et al. 2008). The second well-known factor influencing accuracy of genomic selection prediction is the type of markers used. Solberg et al. (2008) showed that marker type and density determine the accuracy of predictions. In a simulated population, these authors observed that SNP density has to be two to three times greater than microsatellite density to achieve comparable accuracy. Poland (2013) used a recent genotyping-by-sequencing (GBS) method in a set of 254 advanced wheat breeding lines. They compared the accuracy of genomic selection using these markers with the accuracy obtained with DArT markers at an equivalent density. Their results showed that GBS gives more accurate results than DArT. Despite the relative extensive LD in sugarcane, we believe that our marker coverage (1,499 DArT markers) is not sufficient to densely cover the large sugarcane genome (about 120 chromosomes) and to apprehend the totality of the haplotype diversity existing in our panel that represent a core sampling of elite germplasm from numerous current breeding programs. According to Raboin et al. (2008), in sugarcane, the minimum number of multi-allelic locus-specific markers required to achieve a density of one or two markers every 5 cM should be between 300 and 600. Genomic selection prediction for sugarcane could be improved by increasing marker density using multi-allelic markers like microsatellites or GBS markers. Ultimately, we could expect genomic selection predictions to be improved using pedigree information. The work of Crossa et al. (2010) has shown that pedigree added with molecular information can improve predictions accuracy.

In this study, we used two populations that represent a core sampling of elite germplasm more or less recently bred that was picked up from about 30 different sugarcane breeding programs around the world. Our main purpose was to get an insight into the potential of genomic selection relative to complex traits in the context of sugarcane breeding, before possible future implementations applied to local selection programs. Regarding the modest number of markers so far used and the dominant nature of markers, the accuracy of genomic predictions between our two panels seems already very promising, since the ranges of accuracies are similar to those of several genomic selection experiments published for different plant or animal species. These results represent the first concrete illustration of the potential of genomic selection applications relative to complex traits for sugarcane and also for polyploidy crops. The fact that the panels were evaluated in two distant islands, in several environments in both of them, and during several crop cycles for the majority of the traits gives credit to the potential value of genomic selection for concrete perspectives in a particular sugarcane breeding program. The next step will be to experience the efficiency of genomic selection applied to a breeding program. Therefore, it would be interesting to test the performance of genomic selection applied to the hardly selected populations which are encountered in the first stages of breeding programs.

Acknowledgments The authors wish to thank T. Dumont, C. Lallemand, I. Promi, R. Tibère, M. Carbel, J. M. Coupan, O. Calvados and N. Lubin for field work, M. Hoarau for lab work. This study was funded by the eRcane company, by CIRAD (Centre de Coopération Internationale en Recherche Agronomique pour le Développement) ATP-SEPANG project grant, by the Conseil Régional de la Réunion, by the European Union (European regional development fund—ERDF), by ANR (Agence Nationale de la Recherche) Delicas project grant ANR-08-GENM-001, ANR Grass biofuel project grant ANR-07-GPLA-018-005 and by the ANRT (Association Nationale de la Recherche et de la Technologie) through the CIFRE Ph.D grant N°600/2012 of M. Gouy.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards The authors declare that the experiments presented in this publication comply with current French laws.

References

Aitken K, Jackson P, McIntyre C (2006) Quantitative trait loci identified for sugar related traits in a sugarcane (*Saccharum* spp.) cultivar × *Saccharum officinarum* population. Theor Appl Genet 112:1306–1317. doi:10.1007/s00122-006-0233-2

Aitken K, Hermann S, Karno K, Bonnett G, McIntyre L, Jackson P (2008) Genetic control of yield related stalk traits in sugarcane. Theor Appl Genet 117:1191–1203. doi:10.1007/s00122-008-0856-6

Aljanabi S, Parmessur Y, Kross H, Dhayan S, Saumtally S, Ramdoyal K, Autrey L, Dookun-Saumtally A (2007) Identification of a major quantitative trait locus (QTL) for yellow spot (*Mycovellosiella koepkei*) disease resistance in sugarcane. Mol Breeding 19:1–14. doi:10.1007/s11032-006-9008-3

Aufrère J, Baumont R, Delaby L, Peccatte JR, Andrieu J, Andrieu JP, Dulphy JP (2007) Laboratory prediction of forage digestibility by the pepsin-cellulase method. The renewed équations. INRA Prod Anim 20:129–136

Barrière Y, Méchin V, Denoue D, Bauland C, Laborde J (2010) QTL for yield, earliness, and cell wall quality traits in topcross experiments of the F83 × F286 early maize RIL progeny. Crop Sci 50:1761–1772. doi:10.2135/cropsci2009.11.0671

Bates D, Maechler M, Bolker (2011) lme4: linear mixed-effects models using S4 classes. R package version 0.999375-42. http://CRAN.R-project.org/package=lme4. Accessed 11 Oct 2011

Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–561. doi:10.1534/genetics.107.080838

Cheavegatti-Gianotto A, de Abreu H, Arruda P, Bespalhok Filho J, Burnquist W, Creste S, di Ciero L, Ferro J, de Oliveira Figueira A, de Sousa Filgueiras T, Grossi-de-Sá M, Guzzo E, Hoffmann H, de Andrade Landell M, Macedo N, Matsuoka S, de Castro Reinach F, Romano E, da Silva W, de Castro Silva Filho M, César Ulian E (2011) Sugarcane (*Saccharum* × *officinarum*): a reference study for the regulation of genetically modified cultivars in Brazil. Trop Plant Biol 4:62–89. doi:10.1007/s12042-011-9068-3

Collard BCY, Mackill DJ (2008) Marker assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc Lond B Biol Sci 363:557–572. doi:10.1098/rstb.2007.2170

Costet L, Le Cunff L, Royeart S, Raboin LM, Hervouet C, Toubi L, Telismart H, Garsmeur O, Rousselle Y, Pauquet J, Nibouche S, Glaszmann JC, Hoarau JY, D'Hont A (2012a) Haplotype structure around *Bru*1 reveals a narrow genetic basis for brown rust resistance in modern sugarcane cultivars. Theor Appl Genet 125:825–836. doi:10.1007/s00122-012-1875-x

Costet L, Raboin LM, Payet M, D'Hont A, Nibouche S (2012b) A major QTA for resistance to the *sugarcane yellow leaf virus* (Luteoviridae). Plant Breeding 131:637–640. doi:10.1111/j.1439-0523.2012.02003.x

Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. doi:10.1534/genetics.110.118521

Dayal BS, MacGregor JF (1997) Improved PLS algorithms. J Chemom 11:73–85

de los Campos G and Perez R (2010). BLR: bayesian linear regression. R package version 1.2. http://CRAN.R-project.org/package=BLR. Accessed 3 Mar 2012

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345. doi:10.1534/genetics.112.143313

de Roos APW, Hayes BJ, Spelman RJ, Goddard ME (2008) Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. Genetics 179:1503–1512. doi:10.1534/genetics.107.084301

del Blanco IA, Glaz B, Edme SJ (2010) Improving efficiency of sugarcane genotype selection in Florida. Crop Sci 50:1744–1753. doi:10.2135/cropsci2009.09.0539

Denis M, Bouvet JM (2013) Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. Tree Genet Genom 9:37–51. doi:10.1007/s11295-012-0528-1

D'Hont A (2005) Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. Cytogenet Genome Res 109:27–33. doi:10.1159/000082378

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255

FAOSTAT (2012) http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor. Accessed 30 June 2012

Gallais A (1990) Théorie de la sélection en amélioration des plantes, Masson edn. France, Paris, p 588

Glynn NC, Laborde C, Davidson RW, Irey MS, Glaz B, D'Hont A, Comstock JC (2013) Utilization of a major brown rust resistance gene in sugarcane breeding. Mol Breeding 31:323–331. doi:10.1007/s11032-012-9792-x

Grattapaglia D, Resende MDV, Resende MR, Sansaloni CP, Petroli CD, Missiaggia AA, Takahashi EK, Zamprogno KC, Kilian A (2011) Genomic Selection for growth traits in Eucalyptus: accuracy within and across breeding populations. BMC Proc 5(Suppl 7):O16. doi:10.1186/1753-6561-5-S7-O16

Grivet L, Arruda P (2001) Sugarcane genomics: depicting the complex genome of an important tropical crop. Curr Opin Plant Biol 5:122–127. doi:10.1016/S1369-5266(02)00234-0

Guo Z, Tucker D, Lu J, Kishore V, Gay G (2012) Evaluation of genome-wide selection efficiency in maize nested association mapping populations. Theor Appl Genet 124:261–275. doi:10.1007/s00122-011-1702-9

Hadfield JD (2010) MCMC methods for multi–response generalised linear mixed models: the MCMCglmm R package. J Stat Softw 33:1–22

Hadfield JD (2012) MCMCglmm course notes. http://cran.us.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf. Accessed 1 Apr 2013

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12. doi:10.1007/s00122-011-1702-9

Heller-Uszynska K, Uszynski G, Huttner E, Evers M, Carlig J, Caig V, Aitken K, Jackson P, Piperidis G, Cox M, Gilmour R, D'Hont A, Butterfield M, Glaszmann JC, Kilian A (2011) Diversity arrays technology effectively reveals DNA polymorphism in a large and complex genome of sugarcane. Mol Breeding 28:37–55. doi:10.1007/s11032-010-9460-y

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146–160. doi:10.2135/cropsci2011.06.0297

Hoarau M (1969) Sugar cane analysis by hydraulic press method. Int Sugar J 71:328–332

Hoarau JY, Offmann B, D'Hont A, Risterucci AM, Roques D, Glaszmann JC, Grivet L (2001) Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). Part I: genome mapping with AFLP markers. Theor Appl Genet 103:84–97. doi:10.1007/s001220000390

Hoarau JY, Grivet L, Offmann B, Raboin LM, Diorflar JP, Payet J, Hellmann M, D'Hont A, Glaszmann JC (2002) Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). Part II: detection of QTLs for yield components. Theor Appl Genet 105:1027–1037. doi:10.1007/s00122-002-1047-5

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 42:55–67. doi:10.1080/00401706.2000.10485983

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Briefings Funct Genomics 9:166–177. doi:10.1093/bfgp/elq001

Jannoo N, Grivet L, Dookun A, D'Hont A, Glaszmann JC (1999) Linkage disequilibrium among modern sugarcane cultivars. Theor Appl Genet 99:1053–1060. doi:10.1007/s001220051414

Jannoo N, Grivet L, David J, D'Hont A, Glaszmann J (2004) Differential chromosome pairing affinities at meiosis in polyploid sugarcane revealed by molecular markers. Heredity 93:460–467. doi:10.1038/sj.hdy.6800524

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink J (2011) Genomic selection in plant breeding: knowledge and prospects. Adv Agron 110:77–123. doi:10.1016/B978-0-12-385531-2.00002-5

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120:151–161. doi:10.1007/s00122-009-1166-3

Luan T, Woolliams JA, Lien S, Kent M, Svendsen M, Meuwissen THE (2009) The accuracy of genomic selection in norwegian red cattle assessed by cross validation. Genetics 183:1119–1126. doi:10.1534/genetics.109.107391

Meuwissen T, Hayes B, Goddard M (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Mevik BH, Wehrens R, Liland KH (2011) pls: partial least squares and principal component regression. R package version 2.3–0. http://CRAN.R-project.org/package=pls. Accessed 3 Mar 2012

Ming R, Wang Y, Draye X, Moore P, Irvine J, Paterson A (2002) Molecular dissection of complex traits in autopolyploid: mapping QTLs affecting sugar yield and related traits in sugarcane. Theor Appl Genet 105:332–345. doi:10.1007/s00122-001-0861-5

Moser G, Tier B, Crump RE, Khatkar MS, Raadsma HW (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol 41:56. doi:10.1186/1297-9686-41-56

Park T, Casella G (2008) The bayesian lasso. J Am Stat Assoc 103:681–686. doi:10.1198/016214508000000337

Pérez P, de los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Genome 3:106–116. doi:10.3835/plantgenome2010.04.0005

Pinto L, Garcia AAF, Pastina M, Teixeira LHM, Bressiani J, Ulian E, Bidoia MAP, Souza A (2010) Analysis of genomic and functional RFLP derived markers associated with sucrose content, fiber and yield QTLs in sugarcane (*Saccharum* spp.) commercial cross. Euphytica 172:313–327. doi:10.1007/s10681-009-9988-2

Piperidis N, Jackson P, D'Hont A, Besse P, Hoarau JY, Courtois B, Aitken K, McIntyre C (2008) Comparative genetics in sugarcane enables structured map enhancement and validation of marker-trait associations. Mol Breeding 21:233–247. doi:10.1007/s11032-007-9124-8

Poland J (2013) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome Head Print. doi:10.3835/plantgenome2012.06.0006 (Published online 10 Sept. 2012)

R Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. doi:http://www.R-project.org/. ISBN 3-900051-07-0

Raboin L, Offmann B, Hoarau J, Notaise J, Costet L, Telismart H, Roques D, Rott P, Glaszmann J, D'Hont A (2001) Undertaking genetic mapping of sugarcane smut resistance. Proceedings of the South African Sugar Technologists Association, Durban, pp 94–98

Raboin L, Hoarau J, Costet L, Telismart H, Glaszmann J, D'Hont A (2003) Progress in genetic mapping of sugarcane smut resistance. Proceedings of the South African Sugar Technologists Association, Durban, pp 134–141

Raboin LM, Pauquet J, Butterfield M, D'Hont A, Glaszmann JC (2008) Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. Theor Appl Genet 116:701–714. doi:10.1007/s00122-007-0703-1

Reffay N, Jackson PA, Aitken KS, Hoarau JY, D'Hont A, Besse P, McIntyre CL (2005) Characterisation of genome regions incorporated from an important wild relative into Australian sugarcane. Mol Breeding 15:367–381. doi:10.1007/s11032-004-7981-y

Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA, Pappas Jr. GJ, Kilian A, Grattapaglia D (2012a) Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 194:116–128. doi:10.1111/j.1469-8137.2011.04038.x

Resende MFR, Munoz P, Resende MDV, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M (2012b) 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

Rutkoski J, Benson J, Jia Y, Brown-Guedira G, Jannink JL, Sorrells M (2012) Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. Plant Genome 5:51–61. doi:10.3835/plantgenome2012.02.0001

Sabatier D, Thuriès L, Bastianelli D, Dardenne P (2012) Rapid prediction of the lignocellulosic compounds of sugarcane biomass by near infrared reflectance spectroscopy: comparing classical and independent cross validation. J Near Infrared Spectrosc 20:371–385. doi:10.1255/jnirs.999

Schenck S, Hu JS, Lockhart BE (1997) Use of a tissue blots immunoassay to determine the distribution of sugarcane yellow leaf virus in Hawaii. Sugar Cane 4:5–8

Schölkopf B, Smola AJ (2002) Learning with kernels: Support vector machines, regularization, optimization, and beyond. Cambridge, MA, p 626

Scortecci K, Creste S, Calsa T, Xavier M, Landell M, Figueira A, Benedito V (2012) Challenges, opportunities and recent advances in sugarcane breeding. In: Abdurakhmonov I (ed) Plant Breeding. InTech, Rijeka, pp 267–296

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and densities. J Anim Sci 86:2447–2454. doi:10.2527/jas.2007-0010

Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer-Verlag, Heidelberg

Van Soest PJ, Robertson JB, Lewis BA (1991) Methods for dietary fibre, neutral detergent fibre and non starch polysaccharides in relation to animal nutrition. J Dairy Sci 74:3583–3597. doi:10.3168/jds.S0022-0302(91)78551-2

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16–24. doi:10.3168/jds.2008-1514

Wei X, Jackson P, McIntyre C, Aitken K, Croft B (2006) Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure. Theor Appl Genet 114:155–164. doi:10.1007/s00122-006-0418-8

Wei X, Jackson PA, Hermann S, Kilian A, Heller-Uszynska K, Deomano E (2010) Simultaneously accounting for population structure, genotype by environment interaction, and spatial variation in marker-trait associations in sugarcane. Genome 53:973–981. doi:10.1139/G10-050

Whittaker JC, Thompson R, Denham MC (2000) Marker assisted selection using ridge regression. Genet Res 75:249–252

Wold S (2001) Personal Memories of the Early PLS Development. Chemom Intell Lab Syst 58:83–84. doi:10.1016/S0169-7439(01)00152-6

Xu Y, Crouch JH (2008) Marker assisted selection in plant breeding: from publications to practice. Crop Sci 48:391–407. doi:10.2135/cropsci2007.04.0191

Zhong S, Dekkers JCM, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182:355–364. doi:10.1534/genetics.108.098277