Mol Breeding (2020) 40: 38
https://doi.org/10.1007/s11032-020-01120-0

Genomic prediction of sugar content and cane yield in sugar cane clones in different stages of selection in a breeding program, with and without pedigree information

Emily Deomano · Phillip Jackson · Xianming Wei · Karen Aitken · Raja Kota · Paulino Pérez-Rodríguez

Received: 6 August 2019 / Accepted: 16 March 2020 / Published online: 30 March 2020

Abstract High cane yield and commercially extractable sucrose (CCS) content are two of the key sugarcane commercial traits selected in sugarcane breeding programs. Advancements in genomic prediction may provide opportunities to speed up gains for these traits in breeding programs by combining accurate prediction of breeding values in candidate parent clones shortening generation intervals. Selection trials in commercial breeding programs may provide training populations for developing genomic predictions. In this study, three different populations of clones in early and advanced stage selection trials in an established commercial sugarcane breeding program were used to assess genomic prediction accuracy. The clones (genotypes) were evaluated for cane yield and sugar content in field trials and genotyped using a SNP array developed for sugarcane cultivars and parents. Five models (Bayes A, Bayes B, Bayesian LASSO, Bayesian GBLUP and RKHS) were tested using pedigree and/or marker data. Prediction models that included marker information had higher prediction accuracies than models with pedigree data only. For CCS, the prediction accuracies for genotypes in advanced stage trials using DNA markers were superior compared with prediction accuracies for early-stage trials, suggesting that prior intensive selection for CCS did not diminish accuracy of genomic prediction. However, by contrast, for cane yield, the prediction accuracies were much less for the population in the advanced stages of selection. The levels of prediction accuracy obtained in most datasets (0.25–0.45) are encouraging for developing applications of genomic prediction to predict breeding values of yield and sugar content in sugarcane breeding programs.

Key message

- Genomic prediction of cane yield and sugar content in populations of sugarcane in early and advanced stages of selection using a range of models provide encouraging levels of accuracies for developing practical applications in sugarcane breeding.

E. Deomano (✉)
Sugar Research Australia, 50 Meiers Road, Indooroopilly, QLD 4068, Australia
e-mail: EDeomano@sugarresearch.com.au

P. Jackson
CSIRO Agriculture and Food, Australian Tropical Science Innovation Precinct, James Cook University, Townsville, QLD 4811, Australia

X. Wei
Sugar Research Australia, 26135 Peak Downs Highway, Te Kowai, QLD 4741, Australia

K. Aitken
CSIRO Agriculture and Food, Queensland Bioscience Precinct, 306 Carmody Road, St Lucia, QLD 4067, Australia

R. Kota
Syngenta Crop Protection LLC, Research Triangle Park, NC, USA

P. Pérez-Rodríguez
Colegio de Postgraduados, Montecillos, Edo. de México 56230, México

Keywords Genomic selection · Marker assisted breeding · Sugarcane breeding · GBLUP · Bayesian LASSO · RHKS

Introduction

Sugarcane is an important crop in many tropical and sub-tropical regions, grown in more than 100 countries. Development of new cultivars through breeding has been a high priority for nearly all sugarcane industries and sugarcane breeding programs in at least 40 countries for over 100 years (http://www.sugarcanevariety.org/). High cane yield (expressed as tonnes of cane per hectare) and commercially extractable sucrose content (expressed as commercial cane sugar, CCS, in % fresh weight of cane) are important for reducing cost of production of sugar per tonne and are therefore two key selection criteria in sugarcane breeding programs (Jackson 2018).

Sugarcane breeding programs have made critical contributions to productivity, particularly in relation to disease resistance and improvement of ratooning performance (i.e. regrowth after harvest) (Heinz 1987). However, reduced rates of improvement of productivity in sugarcane in recent decades in many countries have led to suggestions for major changes in sugarcane breeding systems (e.g. Wei and Jackson 2017). In particular, it seems likely that low narrow sense heritability for cane yield and sugar content in modern breeding populations and long generation intervals (> 8 years) between crossing and parental selection are contributing factors to low rates of genetic gain in cane yield and sugar content in recent decades (Jackson 2018; Rattey et al. 2004).

Modern sugarcane cultivars and parents are complex hybrid polyploids originally derived from several clones of two species: the high sucrose, octaploid *Saccharum officinarum* ($2n = 80$; $x = 10$) and the low sucrose *S. spontaneum* ($2n = 48–128$; $x = 8$) (Arceneaux 1967; Daniels and Roach 1987; Roach 1989). Previous studies have suggested that this combination of relatively few founders in breeding programs and small number of cycles of inter-crossing (< 9) leads to considerable linkage disequilibrium persisting in sugarcane breeding programs, making marker-assisted selection and breeding approaches feasible despite the large genome and genetic complexity of this crop (Jannoo et al. 1999; Raboin et al. 2008; Wei et al. 2010).

Genomic prediction and selection is an approach increasingly applied in animal and plant breeding programs. Genomic prediction commonly refers to methodologies where all markers are used simultaneously in a statistical model to predict genetic or breeding values of candidate genotypes. This contrasts with other marker-based strategies which focus mostly on individual markers or QTL with strong statistical evidence for significant phenotypic effects. An extensive number of studies have been published on assessing the prediction accuracy of genomic selection in different crops (e.g. Crossa et al. 2013; Pérez-Rodríguez et al. 2012). The accuracy of genomic prediction may be affected by many factors, including statistical models used, the size of the training data, marker density, the heritability of the trait, degree of linkage disequilibrium and the underlying genetic architecture of the traits of interest. Practical applications of genomic selection have been led by studies in dairy cattle and have now been widely reported in animals and crops (e.g. Hayes et al. 2009a; Heffner et al. 2009; Matei et al. 2018; Meuwissen et al. 2001; Vélez-Torres et al. 2018). Most of the benefits of genomic selection arise from obtaining accurate predictions in early stages of the breeding cycle and shorter breeding cycles.

Gouy et al. (2013) provided the first reported study of genomic selection in sugarcane. They screened two separate panels of 167 sugarcane clones with 1499 DArT markers. They found small to moderate levels of accuracy (0.11 to 0.62), with large variation between different traits. The study was encouraging with respect to the potential for practical application in sugarcane breeding, given the accuracy levels in some traits, and some acknowledged limits of the study such as the relatively small number of markers in relation to the large genome size of sugarcane and the relatively small training population size.

In developing genomic prediction models for application in breeding programs, large numbers (e.g. > 1000) of genotypes are desirable for developing accurate genomic prediction models (Hayes et al. 2009b; Van Grevenhof and Van Der Werf 2015). However, collection of accurate phenotypic data is often difficult and expensive particularly for large plants like sugarcane. One option is to utilise data already being collected within commercial breeding programs (Poland 2015; Rutkoski et al. 2015). However, several issues need careful consideration with this approach. First, there are possible advantages and disadvantages in using data from different stages of the selection process. In the early stages of selection, there are usually large numbers of genotypes which have not yet undergone intense selection pressure. In these populations, relatively high genetic variation for most traits typically exist, and this is desirable for training a robust model and developing accurate predictions. However, usually only small plots are used and therefore experimental error variation and inter-plot competition effects are typically large, especially for cane yield in the case of sugarcane (Jackson and McRae 2001). By contrast, in the later stages of selection, most poorly performing clones have already been discarded (in the earlier stages of selection), but measurements are more precise with larger plots, with greater replication.

In this paper, we report the results of a study to assess genomic selection in several populations of sugarcane clones generated in a commercial sugarcane breeding program for the key traits cane yield and commercially extractable sugar content. This report adds to the study of Gouy et al. (2013) in several respects: (i) it includes the important commercial trait of cane yield (not studied by Gouy et al. 2013), (ii) it assesses to what extent DNA markers increase prediction accuracies compared with that possible from using pedigree data, (iii) it compares prediction accuracies using panels of genotypes representing either early or later stages of selection in breeding programs and (iv) it uses a larger number of SNP markers. The results are discussed in relation to the application of genomic selection models in sugarcane breeding programs.

Materials and methods

Genetic material and collection of phenotypic data

Three different panels of sugarcane clones were used. All panels represented genetic material derived from and undergoing selection in the Australian sugarcane breeding program of Sugar Research Australia (SRA). The Australian sugarcane breeding program is similar to most other sugarcane breeding programs worldwide in that most parental clones trace back to a limited number (approximately 20) key ancestral clones generated in breeding programs in Indonesia and India in the early 1900s (Roach 1989).

The three panels of genotypes and the associated phenotypic data collected on these were as follows:

(i). Panel 1 comprised 467 clones used in a previously reported association mapping study (Wei et al. 2010). These clones and the phenotypic data collected were described in detail by Wei et al. (2010). In brief, about half of the genotypes in this panel comprised commercially grown cultivars and/or key parents typically used in the Australian sugarcane breeding program, and the other half were unselected clones from 30 crosses (eight clones per cross) derived from such cultivars and parents. Field trials were established at three locations (Kalamia, Brandon, Victoria mill) in 2006 in Northern Queensland, Australia. The first two were in the Burdekin region, and the latter in the Herbert region, both being important sugarcane producing regions within Australia. Approximately 80% of the 467 genotypes were in common across all three sites with the other clones planted at two of the three sites. At each site, a randomised complete block design, with two replicates per clone, was used. The unit plot size was one row × 10 m, with an inter-row spacing of 1.5 m. Cane yield was measured at approximately 12 months after planting at each site by mechanical harvesting and weighing, and CCS was measured as a random three-stalk sample of stalks taken at the same time at harvest and using procedures routinely used in the Australian sugarcane breeding program (BSES 1984).

(ii). Panel 2 comprised a total of 1146 clones planted in a series of four field trials planted in 2013, with one trial in each of the four major sugarcane growing regions along the Australian east coast at sites Meringa, Burdekin, Mackay and Bundaberg. The experimental design consisted of 1 row × 10 m long plots and the p-design (Cullis et al. 2006) with 8–14% test clones replicated. The number of clones tested in each region varied but this study only used those clones common to all sites. Cane yield was estimated from the weight of the 10 m plot and CCS by small mill method (BSES 1984) or Spectra-Cane (Berding and Marston 2010) based on a sample of five randomly selected stalks. Apart from the trial in Burdekin, data were collected from two crops in each trial: plant crop harvested in 2014 and the first ratoon crop in 2015.

(iii). Panel 3 comprised a total of 738 clones planted in a series of trials in four regions in 2013. In each region, the trials were planted in four to five different locations (giving a total of 18 trials) as final assessment trials of the SRA breeding program. Each individual trial consisted of a p-design with 4-row × 10 m plots with about 15% of clones being replicated. The trials were harvested in 2014, from which cane yield was estimated from the middle two plots and CCS by the press method (BSES 1984) or Spectra-Cane based on six randomly selected stalks.

Analysis of phenotypic data

A general linear mixed model was fitted to the data with spatial variation effects using ASREML-R v3 (Butler et al. 2009). Data analyses for each panel and trait combination were carried out in two steps. First, a model which accounted for spatial and competition effects (Stringer et al. 2011) was developed for each site. Spatial effects were accounted for by determining an optimal model within each selection trial following the methodology of Smith et al. (2007). Then, based on the outcomes of the first step, a combined analysis over all sites was conducted. BLUPs for the clones based on the combined analyses were centred for each site and then averaged over all trials and regions before being used for genomic selection. A combined analysis over all trials was carried out within each of the panel datasets above to obtain BLUPs for each clone. It should be noted that a combined analysis for the BLUPs of clonal effects over all three panels in this study was not feasible because of the poor connectedness between the different panels.

Broad sense heritability was estimated for each trial × trait combination from:

$$H_{\mathrm{b}} = {\sigma_{\mathrm{g}}}^2/\left({\sigma_{\mathrm{g}}}^2 + {\sigma_{\mathrm{e}}}^2/r\right)$$

where ${\sigma_{\mathrm{g}}}^2$, ${\sigma_{\mathrm{e}}}^2$ and r are the genetic variance, error variance and number of replicates per clone, respectively, within each trial.

Pedigree data

Pedigree data over three to ten generations for > 90% of clones in each panel was retrieved from information on ancestors in a database owned by SRA. A small proportion of clones (< 10%) lacked information on ancestry. The total number of ancestor clones used was 438, 708, 880 and 1086, for panels 1, 2, 3 and 2 and 3 combined, respectively. The kinship matrix was computed from the pedigree using the R package 'kinship2' (Sinnwell et al. 2014). The kinship matrix was multiplied by two to obtain the numerator relationship matrix, $\boldsymbol{A}$ which gives the covariance between relatives and was used as the pedigree-based relationship matrix in the models discussed below.

Marker data

All clones were genotyped using an Affymetrix Axiom SNP array developed for sugarcane containing 47,803 SNPs. The development of this array was described by Aitken et al. (2016).

High-quality DNA was extracted from all clones using a standard CTAB extraction method followed by a proteinase K digestion and purification on a Qiagen column as recommended by Affymetrix. The Axiom assay was performed on 96-sample Axiom array using the Affymetrix GeneTitan system according to the procedure described by Affymetrix (http://media.affymetrix.com/support/downloads/manuals/axiom_2_assay_auto_workflow_user_guide.pdf). The final sugarcane Axiom array contained 58,364 probe sets representing 47 ,803 SNPs which were highly polymorphic and all clones were screened across this array. Samples that had a dish quality control (DQC) measure of less than 0.82 or a quality control (QC) call rate of less than 97% were excluded from the analysis. Allele calling was preformed using generated CEL files with Axiom Analysis Suite (1.1.0.616) in order to obtain high-quality results the Axiom best practice workflow was used (http://media.affymetrix.com/support/downloads/manuals/axiom_genotyping_solution_analysis_guide.pdf).

Prior to analysis monomorphic SNP markers were removed. For each polymorphic marker, all clones (genotypes) were given a marker score of 1 if only the most frequent allele was present (i.e. homozygous for this allele), 0 if both alleles were present (i.e. heterozygous), and − 1 if only the minor allele was present. Markers in which one of these three classes occurred for > 98% of the clones were deleted. The number of SNP markers remaining for the analysis after applying the described quality controls were 47,531 for panel 1 and 57,668 for both panels 2 and 3 and 57,675 for panels 2 and 3 when panels 2 and 3 were combined. For each marker, missing values were imputed using observed allelic frequencies. Markers were centered and standardised using sample mean and sample standard deviation.

Analysis of marker effects and genomic predictions

The models described below were evaluated to assess the extent to which use of marker data could provide better prediction accuracies than the use of pedigree information alone. Models based on pedigree data only, marker data only and both pedigree data and marker data combined were considered in the analysis, and these are described briefly below.

(i). Model using only pedigree data

The following model was fitted:

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{u} + \boldsymbol{e} \quad (1)$$

where $\boldsymbol{y}$ is the vector of adjusted phenotypes of each clone, μ is the general mean, $\boldsymbol{u}$ is a vector of random effects with the assumption that $\boldsymbol{u}\sim\mathrm{MN}\left(\mathbf{0}, \sigma_a^2\boldsymbol{A}\right)$ with $\boldsymbol{A}$ the additive relationship matrix derived from pedigree, σ_a^2 the variance component associated to $\boldsymbol{A}$, and $\boldsymbol{e}$ is the residual vector, with the assumption that $\boldsymbol{e}\sim\mathrm{MN}\left(\mathbf{0}, \sigma_e^2\boldsymbol{I}\right)$ with σ_e^2 the residual variance and $\boldsymbol{I}$ the identity matrix, $\boldsymbol{u}$ and $\boldsymbol{e}$ are assumed to be distributed statistically independent.

(ii). Models using marker data only

Several methods, namely, Bayes A, Bayes B, Bayesian LASSO and Ridge Regression (equivalent to GBLUP) (e.g. Crossa et al. 2010; Pérez and de los Campos 2014) were used to model the marker data. These methods differ based on assumptions on the prior distribution assigned to marker effects. The methods are based on the model:

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{X\beta} + \boldsymbol{e} \tag{2}$$

where $\boldsymbol{y}$, μ and $\boldsymbol{e}$ were described previously, $\boldsymbol{X}$ is the matrix for markers of dimension $n \times p$ which has been centered and standardised, $\boldsymbol{\beta}$ is the vector of marker effects, $n =$ number of clones and $p =$ number of markers.

In addition to the above methods, the Reproducing Kernel Hilbert Space (RKHS) kernel model (de los Campos et al. 2010) was fitted. This aims to capture non-additive genetic effects and generate variance covariance structures that are known as reproducing kernels for individuals using the markers. There are several reproducing kernels that can be used, for example Gaussian, exponential, *t*, linear etc. (see for example Tusell et al. 2014). In the case of the Gaussian kernel, it is necessary to specify what is known as bandwidth *h*. The kernels $\boldsymbol{K}_1$, $\boldsymbol{K}_2$ and $\boldsymbol{K}_3$ used three different bandwidth parameters, that is $h \in \left\{\frac{5}{q_{0.05}}, \frac{1}{q_{0.05}}, \frac{1}{5}/q_{0.05}\right\}$ (González-Camacho et al. 2012). We applied the method proposed by de los Campos et al. (2010) known as RKHS-Kernel Averaging (RKHS-KA), in which predictions in the case of three kernels are obtained as follows:

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{u}_3 + \boldsymbol{e} \tag{3}$$

where $\boldsymbol{u}_1 \sim \mathrm{MN}\left(\mathbf{0}, \sigma_1^2\boldsymbol{K}_1\right)$, $\boldsymbol{u}_2 \sim \mathrm{MN}\left(\mathbf{0}, \sigma_2^2\boldsymbol{K}_2\right)$ and $\boldsymbol{u}_3 \sim \mathrm{MN}\left(\mathbf{0}, \sigma_3^2\boldsymbol{K}_3\right)$, $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$ independent.

(iii). Models using both markers and pedigree data

The markers and the pedigree can be used jointly in order to perform the prediction. The model used for prediction based on the Ridge Regression, Bayes A and Bayes B methods is given by:

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{X\beta} + \boldsymbol{u} + \boldsymbol{e} \tag{4}$$

where all the terms were already described above. The model used in the case of the RKHS method is given by:

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{u}_1 + \boldsymbol{u}_2 + \boldsymbol{u}_3 + \boldsymbol{u}_4 + \boldsymbol{e} \tag{5}$$

where $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$ where described above and $\boldsymbol{u}_4 \sim \mathrm{MN}\left(\mathbf{0}, \sigma_a^2\boldsymbol{A}\right)$, $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, $\boldsymbol{u}_3$ and $\boldsymbol{u}_4$ distributed independently.

The BGLR R-package (Pérez and de los Campos 2014) was used to fit all models with default values provided by the software. Inferences were based on 30,000 samples with a burn-in period of 10,000 samples.

Assessing model accuracy

The accuracy of each model described above for predicting cane yield and CCS for each panel of genotypes was evaluated using a cross validation procedure. Initially for each cross validation, 80% of the clones in a particular panel were randomly selected and assigned to a training (TRN) dataset. The remaining 20% of clones in the panel were assigned to a test (TST) dataset. We generated 50 random partitions with 80% of clones in the training dataset and the remaining 20% in the testing dataset. The prediction accuracy was evaluated using the Pearson's product-moment correlation coefficient between the observed phenotype and predicted phenotype for the TST dataset. The average accuracy across the 50 runs was then calculated per method per model.

Results

Analysis of phenotypic data

Figure 1 shows summary statistics from the analyses of cane yield and CCS. Summary statistics were used to evaluate the first, second and third panels of genotypes, respectively. In most datasets, mean levels of both cane yield and CCS were within ranges that are normally expected in commercial production in the regions in which the field trials were located. In panels 1 and 2, the genetic coefficients of variation for cane yield were considerably larger than panel 3. This is probably due to (i) the single row plots used in panels 1 and 2 giving rise to inflated estimates of genetic variance due to competition effects and (ii) clones in panel 3 were selected in prior selection stages and most of the very low yielding

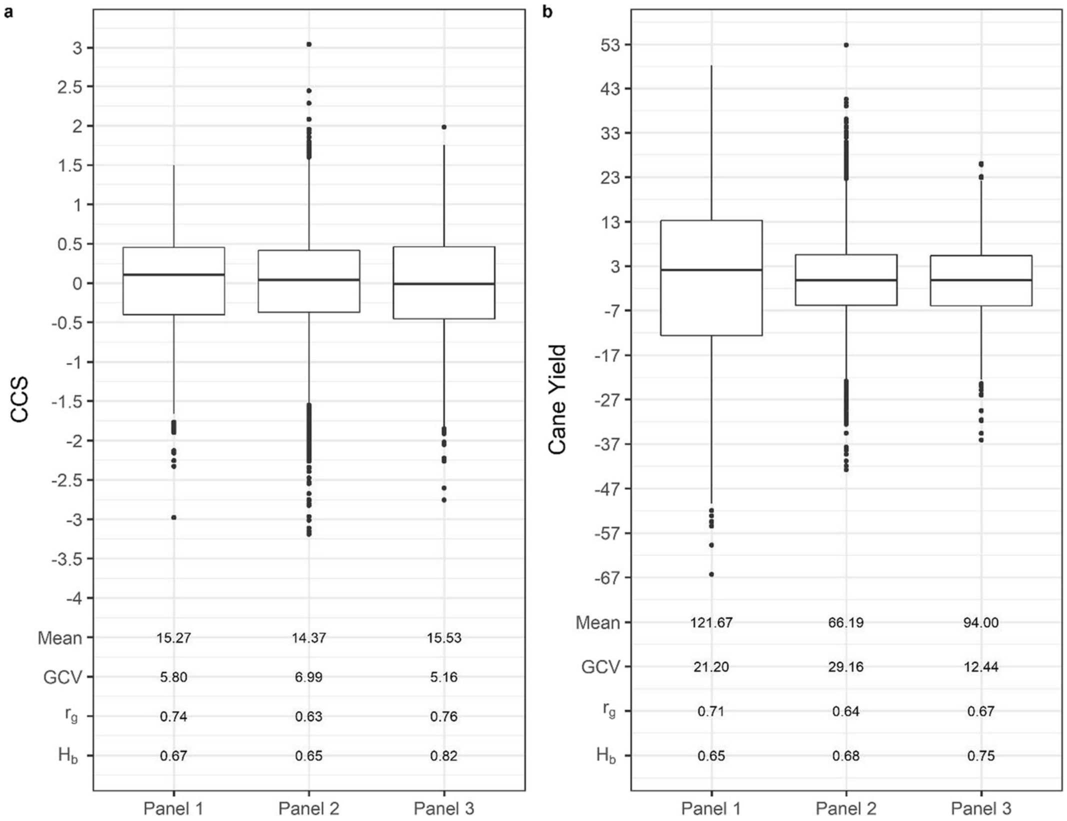

Fig. 1 Box plots showing median BLUPs, average predicted means, average broad sense heritability (H_b), genetic coefficient of heritability (GCV) and average genetic correlation (r_g mean) for panels 1, 2 and 3

clones were probably eliminated in these prior stages. The genetic coefficients of variation for CCS were also slightly larger for panels 1 and 2, but this difference was not as large as for cane yield, probably reflecting the much smaller impact of competition in single row plots on sugar content than cane yield (Jackson and McRae 2001). For all three panels, the broad sense heritability for both cane yield and CCS and the genetic correlations between trials were moderate to high (> 0.6), indicative of strongly repeatable genetic effects and accurate overall discrimination among genotypes. Even though the field trials in panels 1 and 2 comprised smaller plots and less replication than panel 3, the possible impact on broad sense heritability due to greater error variance was probably offset by higher genetic variation. Overall, these phenotypic data for cane yield and CCS may be considered typical of reasonable to good quality data routinely generated in selection trials in the Australian sugarcane breeding program, without any unusual features, and suitable for examining the accuracy of genomic selection approaches.

Accuracy of genomic prediction methods

For the first panel of genotypes for CCS, models using pedigree information alone were similar to models using marker data (Fig. 2). The combination of pedigree and marker data was similar or slightly superior to where marker data were used alone. For cane yield in the first panel, use of marker data did not provide greater accuracy in prediction than the pedigree data alone, and the combination of marker and pedigree data did not provide any superiority over the use of pedigree data alone, and in fact, was slightly inferior in all cases. The Bayes B method gave lower accuracies than the other methods for the models using marker data.

For panel 2, all methods gave similar results except for Bayes B which gave lower accuracies for models including marker data (Figs. 2 and 3). The use of marker data alone was superior to the use of pedigree data alone for both CCS and cane yield, and the joint use of pedigree and marker data provided greater accuracy in the case of cane yield but not for CCS.

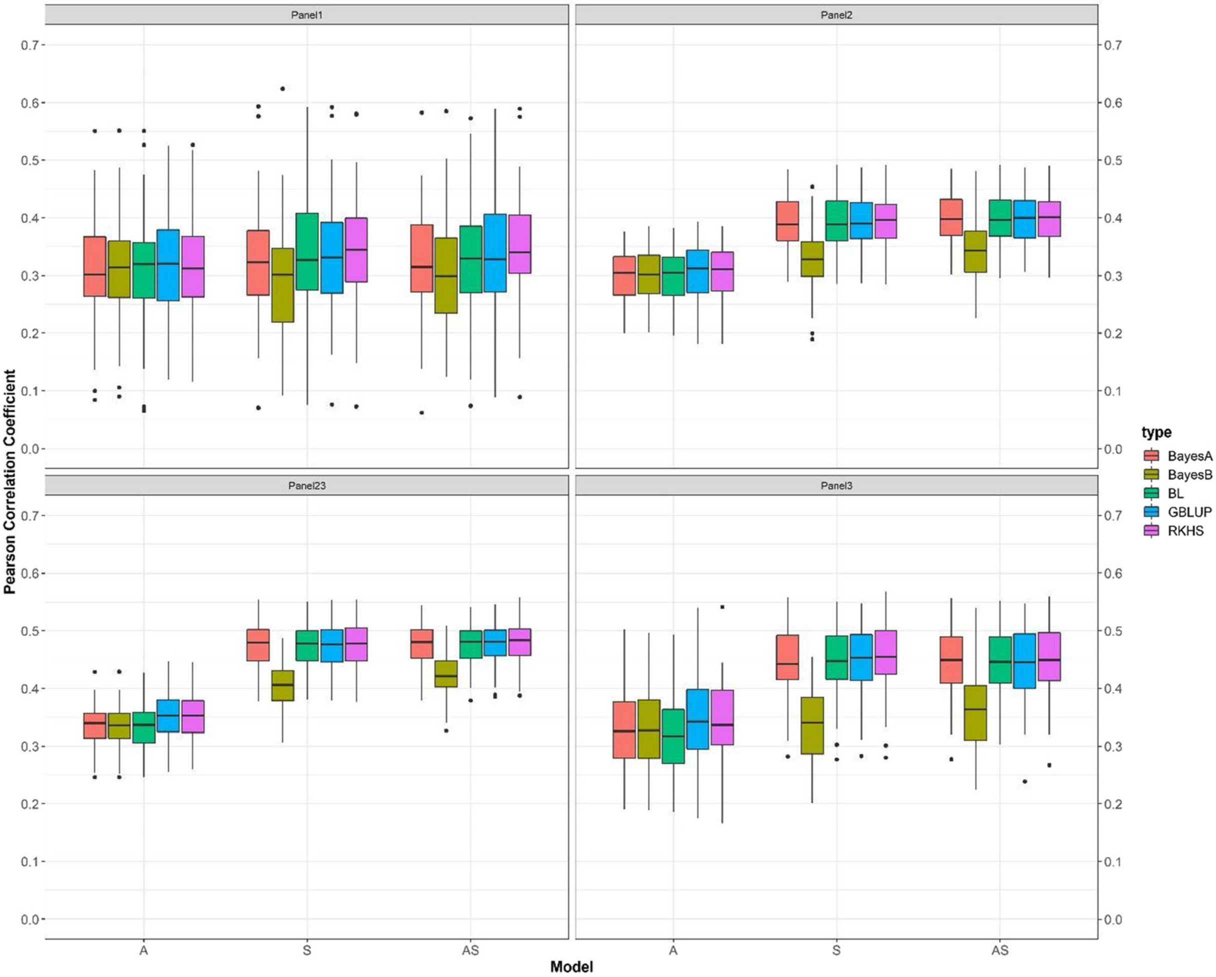

Fig. 2 Box plots showing median accuracies of predictions for commercial cane sugar (CCS) for models using only pedigree data (A), models using only marker data (S) and models using both marker and pedigree data (AS) for panels 1, 2, 2 and 3 combined and 3. For each data × panel combination, accuracies are given (from left to right) for Bayes A, Bayes B, Bayesian Lasso, Ridge regression (GBLUP) and RHKS modelling methods

For panel 3 for CCS and cane yield, all methods gave similar levels of prediction accuracy, except for Bayes B which gave lower accuracy for models using marker data (Figs. 2 and 3). However, accuracy levels for cane yield overall in panel 3 were considerably lower than for the other two panels (Fig. 3). The pedigree data alone gave low levels of prediction accuracy particularly for cane yield compared with the other datasets. By contrast for CCS, all methods provided prediction accuracies based on marker data which were similar or higher than for the other datasets. For cane yield, the combination of marker and pedigree data gave similar levels of accuracy to the model where marker data were used alone while for CCS the combination of marker and pedigree data provided a similar accuracy compared with markers alone.

When panels 2 and 3 were combined, the accuracies obtained were similar to what was obtained for panel 2 in respect to cane yield, and similar to panel 3, in respect to CCS (Figs. 2 and 3). For both cane yield and CCS, the use of marker data gave improved predictions compared with use of pedigree data alone. The joint use of pedigree and marker data gave improved predictions compared with marker data alone for cane yield, but not for CCS.

Discussion

The use of the same set of markers and prediction models applied to different types of panels of genotypes allowed

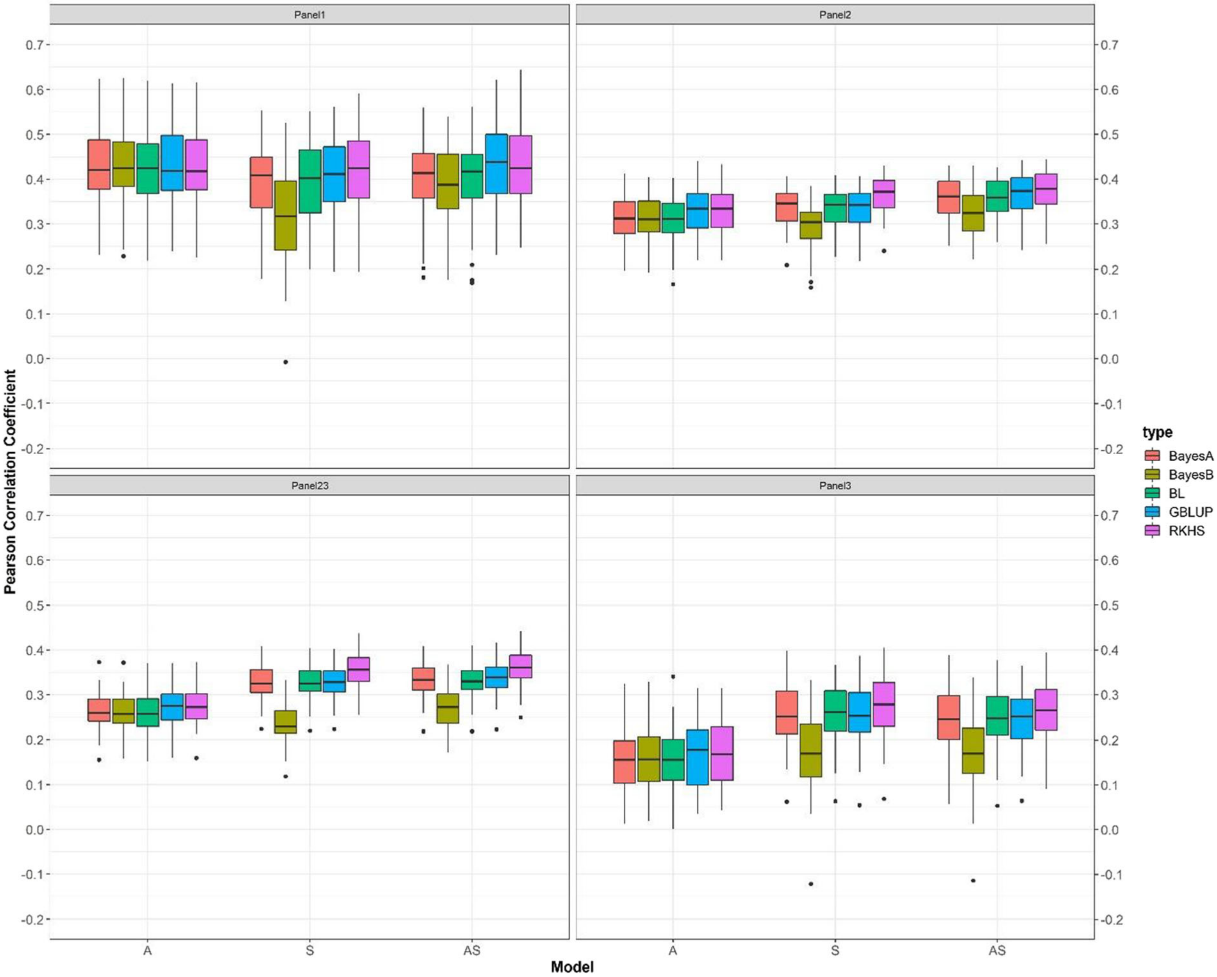

Fig. 3 Box plots showing median accuracies of predictions for cane yield for models using only pedigree data (A), models using only marker data (S) and models using both marker and pedigree data (AS) for panels 1, 2, 2 and 3 combined and 3. For each data × panel combination, accuracies are given (from left to right) for Bayes A, Bayes B, Bayesian Lasso, Ridge regression (GBLUP) and RHKS modelling methods

prediction accuracies to be directly compared across sets of genotypes and phenotypic data with different features. Panels 1 and 2 consisted of relatively less selected clones, and as expected, these appeared phenotypically more diverse compared with panel 3, as indicated by higher genetic coefficients of variation especially for cane yield. Genotypes in panels 1 and 2 were considered more representative of those in early or middle stages of selection in breeding programs, while genotypes in panel 3 were typical of those in advanced stages of selection.

For CCS, the prediction accuracies using DNA markers were superior in panel 3 compared with the other two panels. This indicates that the prior intensive selection for CCS that would have occurred in selecting the clones used in panel 3 did not diminish accuracy of genomic prediction using DNA markers. However, by contrast, for cane yield, the prediction accuracies were much less for panel 3 compared with both panels 1 and 2. There are at least two explanations for this. First, one factor contributing to the reduced genetic variation (as measured by genetic coefficients of variation) in panel 3 was that bordered plots were used in contrast with single-row plots in panels 1 and 2. Cane yield is known to be strongly affected by competition effects in single-row plots, in contrast with CCS which is mostly unaffected (Jackson and McRae 2001). Thus, competition effects would have contributed to the variation in cane yield in panels 1 and 2, and genetic effects for competitive ability may have been included in predictions in panels 1 and 2. Second, prior selection of clones in panel

3, particularly discarding the most inferior genotypes contributing to variation in panels 1 and 2, may have reduced the capacity of the genomic prediction models.

Data from panels 2 and 3 were combined to assess accuracy obtainable from combining data from two different types of selection trials. Combining data from across different stages of selection would be an option for obtaining large training sets of phenotypic data from most breeding programs. In this study, panel 2 was representative of relatively early (stage 2) trials and germplasm in a commercial breeding program, with clones being evaluated in single row plots subject to higher inter-plot competition effects than in later stages. Panel 3, by contrast, represented advanced stages of selection and clones that had undergone prior rounds of effective selection and which were evaluated in large plots mostly free from competition effects. The results indicated prediction accuracies that were slightly superior to those obtained for the other panels. This suggests a strategy of combining data from different stages of selection for creation of large training sets is an acceptable approach.

Resolution of the causes of reduced genomic prediction accuracy for cane yield in panel 3 would seem important for both future research and future practical application of genomic prediction approaches in sugarcane breeding. If the higher accuracies in panels 1 and 2 were largely due to the influence of competition effects, it would mean that genomic predictions of cane yield based on data subjected to competition effects could be unreliable and ineffective for application in breeding programs where the aim is development of cultivars with high yield in pure (commercial) stands, rather than with high competitive ability in mixed plots. If this was the case, genomic prediction models should be confined to panels where the phenotypic data are obtained from multi-row plots. By contrast, if the higher accuracies in panels 1 and 2 were predominately due to the reduced genetic variation in the highly selected clones used for panel 3, then genomic prediction models from data obtained from panels 1 and 2, and other similar early-stage selection trials, would clearly be superior and should be applied. However, resolution of these two causes may not be easy because it would seem to necessitate evaluation of a large population of relatively unselected genotypes in both multi-row and single row plots, and cross validation of genomic predictions derived from the two types of plots. The use of multi-row plots is expensive, and the evaluation of large numbers of unselected genotypes in such plots is not done in commercial breeding programs.

Determining prediction accuracy using marker data with and without pedigree data included in the prediction models was not reported previously for sugarcane in the study by Gouy et al. (2013). However, this is important in helping assess the value of marker data in practice because most sugarcane breeding programs have access to comprehensive pedigree data and can utilise this for predicting breeding values. The additional expense of obtaining marker data in breeding programs is only justified if the use of markers could add to the predictions significantly compared with using the already available pedigree data. In panels 2 and 3, the use of marker data clearly added additional accuracy to the model where pedigree data were used alone, although in panel 1, the marker data did not provide any advantage. The reason for the differences between panels in this respect is not known, and no plausible hypotheses were proposed.

Prediction accuracies reported in this study for CCS were similar to those reported by Gouy et al. (2013) for brix (a measure of soluble solids, which is known to be usually strongly correlated with CCS). Gouy et al. (2013) did not report on cane yield, but accuracies which they found for yield components including stalk number and stalk diameter were of a similar magnitude as we found for cane yield. Gouy et al. (2013) suggested that their accuracy levels in their study may have been limited due to the small number of markers they used (1499). The similar levels of accuracy obtained in our study, despite the much larger number of markers used, is suggestive that the importance of marker number may not be as great as suggested by Gouy et al. (2013). This may be at least in part due to the high level of linkage disequilibrium persisting in sugarcane breeding program populations.

The levels of prediction accuracy obtained in this study, especially for cane yield, were relatively low compared with a range of traits in other crops, particularly when considering the relatively high broad-sense heritability of our observed phenotypes in our study. While highly variable, accuracies reported in a range of traits in other crops, have been higher (e.g. Heffner et al. 2009, 2011; Crossa et al. 2011, 2014; You et al. 2016). One explanation for the low or moderate proportions of genetic variation predicted in our study is that complex non-additive genetic effects, known to comprise a high proportion of genetic variance for a range of traits in sugarcane, and particularly for cane yield (Wei and Jackson 2017), were not explained in the prediction models. The RKHS method has been advocated as having an advantage in being able to model non-additive genetic effects. However, the five different methods of genomic prediction tested in our study gave very similar results in most cases. The fact that the RHKS model had a similar accuracy to the other methods suggests that it may have failed to adequately model complex non-additive genetic effects in these data. Another possible source of variation not accounted for was additive effects due to variation in allele dosages, because the SNP markers were called only as being present or absent, but in a polyploid such as sugarcane variation in numbers of copies of each allele may be important. As discussed below, development of genotyping methods that allow for accurate estimation of marker dosages is a priority for future research.

The genomic prediction models used in our study did not explicitly model genotype x environment interactions, as has been done in some other studies (e.g. Jarquín et al. 2017) for two reasons. Firstly, our goal was to predict average genetic and breeding values across all production environments rather than performance in specific environments, which is in line with the objective of the Australian sugarcane breeding program. Secondly, genotype × environment interactions in sugarcane in Australia are usually smaller than overall genotypic effects, as indicated by reasonably high genetic correlations between selection trials (Fig. 1). Sugarcane genotypes are still tested in selection trials across multiple sites and crop-years in the Australian sugarcane breeding program because this improves the accuracy of estimating overall mean genotypic effects given that genotype × environment interactions are usually significant. However, the relatively small magnitude of these effects, which are often also not repeatable, in comparison with overall genotypic effects, means that there is usually little or no interest by breeders in predicting or exploiting these effects in selection. This situation contrasts with some grain crops and environments, where variation due to genotype × environment interactions may be larger than overall genotype effects. This difference between sugarcane and some other crops may be due to the impact in some crops of variation in time of flowering interacting with some environmental factors such as water stress having major impacts on grain numbers and final yields, while this source of variation is absent in sugarcane.

Marker data used in this study only depicted the presence versus absence of alternative alleles, and not dosage levels of alleles. Because variation in allele dosage is likely to affect trait performance to some extent this is expected to be an important limitation to models used in this study. Methodology to estimate allele dosage has been developed for SNP arrays in autotetraploids (e.g. Schmitz Carley et al. 2017). Methodology to undertake this in higher-level polyploids such as sugarcane has been developed for analysis of genotyping by sequencing data (e.g. Serang et al. 2012; Garcia et al. 2013; Gerard et al. 2018; Mollinari and Garcia 2019). This is also under active investigation for SNP array data, and although progress in this area is expected to occur in coming years, there are challenges in calling accurate dosage levels using SNP array data in higher-level polyploids (You et al. 2018). The estimation of allele dosages and the incorporation of this information in genomic prediction models is an area of priority for ongoing research and in developing practical breeding applications to breeding in sugarcane (as with other higher-level polyploids) in the future.

Two main applications of genomic prediction are being explored in plant breeding programs (Crossa et al. 2014). Firstly, predicting total genetic value of individual genotypes for possible release as cultivars. In this case, prediction of both additive and non-additive effects is important. The other application is focused on selection of parents and rapid cycling of generations. This depends mainly on only prediction of additive effects because these are the main component of total genetic variation which is passed from parents to offspring (Falconer and Mackey 1996). Our results suggest that current models for genomic prediction models are unlikely to offer practical value in sugarcane breeding programs for prediction of genetic value, which is required for selecting clones for commercial release. Large improvements in prediction accuracy of genetic value may only arise for models which effectively capture non-additive genetic effects, given the likely high importance of these effects.

However, application of models which use predominately additive marker effects may have particular value for predicting breeding value in sugarcane breeding programs. Genomic prediction accuracies obtained in our study based on additive effects may correspond to high accuracies for underlying additive genetic effects, given that these effects comprise only a small proportion of the total phenotypic variation being predicted. Predicting breeding values and speeding up gains from selection through shortening the generation interval could be a major contribution of genomic prediction. It has been hypothesised that apparently limited gains in sugarcane productivity and particularly in parental breeding values could be at least partly attributable to the high proportion of genetic variation present as non-additive variation in sugarcane and long generation interval between crossing and parent selection (Wei and Jackson 2017). Even though selection for favourable non-additive variance is obviously important for selection of clones for cultivars, the high non-additive genetic variation and low narrow sense heritability clones in modern sugarcane breeding programs means that selection based on phenotype poorly predicts breeding value. It is possible that genomic prediction based on predominately additive effects of DNA markers may better predict breeding value than use of phenotypic data, and some evidence for this has been reported (Wei and Jackson 2017). This approach is not dependent on methodology accounting for non-additive genetic effects and is therefore perhaps more immediately available for application than applications relying on prediction of total genetic effects.

Author contribution statement XW and PJ initiated the study and coordinated collection of phenotypic data. XW conducted analyses of phenotypic data. KA and RK coordinated development of the SNP array and collation of DNA marker data. ED conducted all analyses related to genomic prediction. PPR assisted in the development of R scripts for fitting Bayesian models. ED, PJ and XW wrote the manuscript, and all authors read the manuscript.

Funding information Financial support for the research reported on here, including the collection of the large datasets, was provided by Sugar Research Australia, Syngenta and Commonwealth Scientific and Industrial Research Organisation. A large number of staff from breeding programs operated Sugar Research Australia and Wilmar Sugar Australia contributed to collecting phenotypic data from field trials, and we are grateful for their competent conduct of these trials.

Compliance with ethical standards

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

References

Aitken K, Farmer A, Berkman P, Muller C, Wei X, Demano E, Jackson P, Magwire M, Dietrich B, Kota R (2016) Generation of a 345K sugarcane SNP chip. Proc Aust Soc Sugar Cane Technol:1165–1172

Arceneaux G (1967) Cultivated sugarcanes of the world and their botanical derivation. Proc Int Soc Sugar Cane Technol:844–854

Berding N, Marston DH (2010) Operational validation of the efficacy of SpectraCane™, a high-spped analytical system for sugarcane qulaity components. Proc Int Soc Sugar Cane Technol 32:445–459

BSES (1984) The standard laboratory manual for Australian sugar Mills. Bureau of Sugar Experiment Stations, Brisbane

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) Mixed models for S language environments ASReml-R reference manual. The Department of Primary Industres and Fisheries. Queensland Government, Australia

Crossa J, de los Campos G, Pérez P (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724

Crossa J, Pérez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokosho C (2011) Genomic selection and prediction in plant breeding. J Crop Improv 25:239–261

Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 Genes Genomes Genetics 3: 1903–1926

Crossa J, Perez P, Hickey J, Burgueno J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 112:48

Cullis B, Smith A, Coombes N (2006) On the design of early generation variety trials. J Agric Biol Environ Stat 11:381–393

Daniels J, Roach BT (1987) Taxonomy and evolution. Sugarcane improvement through breeding, vol 7. Elsevier, Amsterdam, pp 7–84

de los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res 92(4):295–308

Falconer DS and Mackey TFC (1996) Introduction to quantitative genetics, 4th ed. Longman, Burnt Mill, England

Garcia AAF, Mollinari M, Marconi TG, Serang OR, Silva RR, Vieira MLC, Vicentini R, Costa EA, Mancini MC, Garcia MOS, Pastina MM, Gazaffi R, Martins ERF, Dahmer N, Sforça DA, Silva CBC, Bundock P, Henry RJ, Souza GM, van Sluys MA, Landell MGA, Carneiro MS, Vincentz MAG, Pinto LR, Vencovsky R, Souza AP (2013) SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids. Sci Rep 3:3399. https://doi.org/10.1038/srep03399

Gerard D, Ferrão LFV, Garcia AAF, Stephens M (2018) Genotyping polyploids from messy sequencing data. Genetics 210(3):789–807. https://doi.org/10.1534/genetics.118.301468

González-Camacho JM, de los Campos G, Pérez-Rodríguez P, Gianola D, Cairns JE, Mahuku G, Babu R, Crossa J (2012) Genome-enabled prediction of genetic values using radial basis function. Theor Appl Genet 125:759–771

Gouy M, Rousselle Y, Bastianelli D, Lecomte P, Bonnal L, Roques D, Efile JC, Rocher S, Daugrois J, Toubi L, Nabeneza S, Hervouet C, Telismart H, Denis M, Thong-Chane A, Glaszmann JC, Hoarau JY, Nibouche S, Costet L (2013) Experimental assessment of the accuracy of genomic selection in sugarcane. Theor Appl Genet 126:2575–2586

Hayes BJ, Bowman PJ, Chamberlain A, Goddard M (2009a) Invited review: genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92:433–443

Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME (2009b) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41:51

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

Heffner EL, Jannink J-L, Iwata H, Souza E, Sorrells ME (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci 51:2597–2606

Heinz DJ (1987) Sugarcane improvement through breeding. Elsevier

Jackson P (2018) Advances in conventional sugarcane breeding. In: Rott P (ed) Achieving sustainable cultivation of sugarcane volume 2: breeding, pests and diseases. Burleigh Dodds Science Publishing, Cambridge

Jackson P, McRae T (2001) Selection of sugarcane clones in small plots. Crop Sci 41:315–322

Jannoo N, Grivet L, Seguin M, Paulet F, Domaingue R, Rao P, Dookun A, d'Hont A, Glaszmann J (1999) Molecular investigation of the genetic base of sugarcane cultivars. Theor Appl Genet 99:171–184

Jarquín D, Lemes da Silva C, Gaynor RC, Poland J, Fritz A, Howard R, Battenfield S, Crossa J (2017) Increasing genomic-enabled prediction accuracy by modeling genotype by environment interactions in Kansas wheat. Plant Genome 10. https://doi.org/10.3835/plantgenome2016.12.0130

Matei G, Woyann LG, Milioli AS, de Bem Oliveira I, Zdziarski AD, Zanella R, Coelho ASG, Finatto T, Benin G (2018) Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. Mol Breed 38:117

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

Mollinari M, Garcia AAF (2019) Linkage analysis and haplotype phasing in experimental autopolyploid populations with high ploidy level using hidden Markov models. G3 Genes Genomes Genetics 9(10):3297–3314. https://doi.org/10.1534/g3.119.400378

Pérez P, de los Campos G (2014) Genome-wide regression & prediction with the BGLR statistical package. Genetics 114: 164442

Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y, Dreisigacker S (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3: Genes, Genomes, Genetics 2:1595–1605

Poland J (2015) Breeding-assisted genomics. Curr Opin Plant Biol 24:119–124

Raboin L-M, Pauquet J, Butterfield M, D'Hont A, Glaszmann J-C (2008) Analysis of genome-wide linkage disequilibrium in the highly polyploid sugarcane. Theor Appl Genet 116:701–714

Rattey A, Jackson P, Wei X, Cox M, Stringer J (2004) Opportunities to increase rates of parent improvement in Australian sugarcane breeding programs. In: Proceedings—Australian Society of Sugar Cane Technologists. PK Editorial Services, pp 42–42

Roach B (1989) Origin and improvement of the genetic base of sugarcane. Proc Aust Soc Sugar Cane Technol:34–47

Rutkoski J, Singh R, Huerta-Espino J, Bhavani S, Poland J, Jannink J, Sorrells M (2015) Efficient use of historical data for genomic selection: a case study of stem rust resistance in wheat. The Plant Genome 8

Schmitz Carley CA, Coombs JJ, Douches DS, Bethke PC, Palta JP, Novy RG, Endelman JB (2017) Automated tetraploid genotype calling by hierarchical clustering. Theor Appl Genet 130:717–726. https://doi.org/10.1007/s00122-016-2845-5

Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. PLoS One 7(2):e30906. https://doi.org/10.1371/journal.pone.0030906

Sinnwell JP, Therneau TM, Schaid DJ (2014) The kinship2 R package for pedigree data. Hum Hered 78:91–93

Smith AB, Stringer J, Wei X, Cullis BR (2007) Varietal selection for perennial crops where data relate to multiple harvests from a series of field trials. Euphytica 157:253–266

Stringer J, Cullis B, Thompson R (2011) Joint modeling of spatial variability and within-row interplot competition to increase the efficiency of plant improvement. J Agric Biol Environ Stat:1–13

Tusell L, Pérez-Rodríguez P, Forni S, Gianola D (2014) Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. J Anim Breed Genet 131(2):105–115

Van Grevenhof IE, Van Der Werf JH (2015) Design of reference populations for genomic selection in crossbreeding programs. Genet Sel Evol 47:14

Vélez-Torres M, García-Zavala JJ, Hernández-Rodríguez M, Lobato-Ortiz R, López-Reynoso JJ, Benítez-Riquelme I, Mejía-Contreras JA, Esquivel-Esquivel G, Molina-Galán JD, Pérez-Rodríguez P (2018) Genomic prediction of the general combining ability of maize lines (Zea mays L.) and the performance of their single crosses. Plant Breed 137:379–387

Wei X, Jackson P (2017) Addressing slow rates of long-term genetic gain in sugarcane. Int Sugar J 119:480–484

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 this article is one of a selection of papers from the conference "exploiting genome-wide association in oilseed brassicas: a model for genetic improvement of major OECD crops for sustainable farming". Genome 53:973–981

You FM, Booker HM, Duguid SD, Jia G, Cloutier S (2016) Accuracy of genomic selection in biparental populations of flax (Linum usitatissimum L.). Crop J 4:290–303

You Q, Yang X-P, Peng Z, Xu L-P, Wang J-P (2018) Development and applications of a high throughput genotyping tool for polyploid crops: single nucleotide polymorphism (SNP) array. Front Plant Sci 9. https://doi.org/10.3389/fpls.2018.00104

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.