Evaluation of genomic selection and marker-assisted selection in *Miscanthus* and energycane

Marcus O. Olatoye · Lindsay V. Clark ·
Jianping Wang · Xiping Yang · Toshihiko Yamada ·
Erik J. Sacks · Alexander E. Lipka

Received: 4 January 2019 / Accepted: 26 November 2019 / Published online: 7 December 2019
© The Author(s) 2019

Abstract Although energycane (*Saccharum spp.* hybrids) is widely used as a source of lignocellulosic biomass for bioethanol, breeding this crop for disease resistance is challenging due to its narrow genetic base. Therefore, efforts are underway to introgress novel sources of genetic resistance from *Miscanthus* into energycane. Given that disease resistance in energycane could be either qualitative or quantitative in nature, careful examination of a wide variety of genomic-enabled breeding approaches will be crucial to the success of such an undertaking. Here we examined the efficiency of both genomic selection (GS) and marker-assisted selection (MAS) for traits simulated under different genetic architectures in F₁ and BC₁ populations of *Miscanthus* × *Miscanthus* and sugarcane × sugarcane crosses. We observed that the performance of MAS was comparable and sometimes superior to GS for traits simulated with four quantitative trait nucleotides

(QTNs). In contrast, as the number of simulated QTN increased, all four GS models that were evaluated tended to outperform MAS, select more phenotypically optimal F₁ individuals, and accurately predict simulated trait values in subsequent BC₁ generations. We therefore conclude that GS is preferable to MAS for introgressing genetic sources of horizontal disease resistance from *Miscanthus* to energycane, while MAS remains a suitable option for introgressing vertical disease resistance.

Keywords Bioenergy crops · Genomic selection · Marker-assisted selection · Genetic mechanism

Introduction

Dedicated perennial lignocellulosic biomass crops can contribute substantially to meeting society's energy needs while sequestering carbon, building soils, mitigating climate change, and improving livelihoods in rural communities (Saha et al. 2013). Energycane (*Saccharum spp.* hybrids), a type of sugarcane that has been bred for use as a bioenergy feedstock, has great potential as a source of lignocellulose and/or sugar for sustainably producing biofuels (Kandel et al. 2018). Nevertheless, the cultivation and productivity of energycane has been hampered by its narrow genetic base and disease susceptibility because most modern energycane varieties were derived from few interspecific hybrids between *S. officinarum* L. and *S. spontaneum* L. (Wang et al. 2013). To improve disease resistance in energycane, two other genera (*Miscanthus* and

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11032-019-1081-5>) contains supplementary material, which is available to authorized users.

M. O. Olatoye · L. V. Clark · E. J. Sacks · A. E. Lipka (✉)
Department of Crop Sciences, University of Illinois,
Urbana, IL, USA
e-mail: alipka@illinois.edu

J. Wang · X. Yang
Agronomy Department, University of Florida,
Gainesville, FL, USA

T. Yamada
Hokkaido University, Sapporo, Hokkaido, Japan

Erianthus) that share a close phylogenetic relationship with energycane have received appreciable attention (Tew and Cobill 2008). For instance, intergeneric crosses between *S. officinarum* ‘Ludashi’ and *Erianthus rockii* have been used to introgress brown rust resistance into energycane (Wang et al. 2013). Miscane, a hybrid of *Saccharum* $\times$ *Miscanthus*, has been noted to exhibit increased cold tolerance and disease resistance compared with ordinary *Saccharum* and therefore can potentially facilitate disease resistance introgression back into energycane through backcrossing (Chen and Lo 1989). However, breeding processes can be challenging in energycane. The polyploid genome of energycane complicates trait inheritance analysis (Gouy et al. 2013), and thus energycane breeding is still largely dependent upon phenotypic selection. Additionally, large trials and multiple selection cycles (7–10 years) are needed for improvement of traits such as yield and disease resistance (Gouy et al. 2013; Kandel et al. 2018). Therefore, the development and refinement of genomic-enabled breeding methods specifically for energycane has potential to substantially improve the efficiency of breeding for this bioenergy crop.

Markers tagging quantitative trait loci (QTL) and other genomic signals exhibiting statistically significant associations with traits have been identified in energycane (Yang et al. 2018), which can potentially facilitate marker-assisted selection (MAS) in energycane breeding programs. Given the success of MAS breeding programs in other crops (Okogbenin et al. 2007; Xu and Crouch 2008; Jiang 2013; Yohannes et al. 2015), the potential of MAS to develop disease-resistant energycane crops is promising. Nevertheless, MAS-based breeding approaches are undermined by several factors. Most importantly, QTL analyses often miss small-effect loci, and thus MAS is typically based on only a few large-effect loci that may not capture all genetic variation responsible for the trait (Jannink et al. 2010; Ben-Ari and Lavi 2012). Additionally, because mapping populations employed in identifying QTLs used for MAS typically consist of a relatively small number of individuals ($n < 250$), these estimated QTL effects are usually inflated and potentially not reflective of their true genetic effects (Beavis 1995, 1998; Utz et al. 2000; Xu 2003). Given these setbacks, it is not surprising that MAS has been most successfully applied to traits controlled by few genomic loci (Bernardo 2001; Zhao et al. 2014). Currently, QTL in sugarcane and *Miscanthus* has only been able to tag

relatively few large-effect genes for a given trait (Costet et al. 2012; Dong et al. 2018). Thus, the contributions of many small-effect loci underlying horizontal disease-resistance may be unaccounted for in MAS-based breeding programs.

Another approach that is arguably better suited for breeding complex traits such as horizontal resistance is genomic selection (GS), which uses genome-wide marker sets to predict breeding values (Meuwissen et al. 2001). The potential of GS to accelerate the breeding cycle, increase genetic gain, and maintain genetic diversity above what can be achieved through MAS has been demonstrated (Meuwissen et al. 2001; Heslot et al. 2012; Annicchiarico et al. 2015). Previous investigations into the prospects of GS in energycane for traits related to sugar and bagasse contents, plant morphology, and disease resistance have produced promising results, with prediction accuracies of up to 0.62 being reported (Gouy et al. 2013; de Almeida Costa 2015). Moreover, GS has been successfully used for the improvement of complex traits in other crops (Wang et al.; Heffner et al. 2011; Heslot et al. 2012; Resende et al. 2012; Beyene et al. 2015; Gezan et al. 2017), including quantitative disease resistance in maize (Technow et al. 2013) and wheat (Rutkoski et al. 2012, 2015; Mirdita et al. 2015). Therefore, GS has the potential to substantially improve qualitative and quantitative resistance in energycane, and evaluation of its ability to accelerate the development of genetically diverse, disease-resistant cultivars is needed.

Given recent reductions in genotyping costs (Elshire et al. 2011; Poland 2015), the adoption of GS into crop breeding programs is becoming increasingly possible. Although a benefit of these cost reductions is an increased number of available genome-wide markers, several studies have shown that randomly selected subsets of genome-wide markers are capable of achieving similar prediction accuracies relative to the full marker sets (Arruda et al. 2015; Zhang et al. 2015; Spindel et al. 2015). Thus, to ensure that resources for GS-based breeding programs are allocated as effectively as possible, evaluation of the impact of the number of markers (i.e., marker density) on GS prediction accuracy is crucial.

The comparison of GS to MAS has been the subject of previous studies in various crop species, including maize (Massman et al. 2013; Owens et al. 2014; Cao et al. 2017; Cerrudo et al. 2018), wheat (Arruda et al.

2015), rye (Wang et al. 2014), and cowpea (Olatoye et al. 2019). The results of this prior work was consistent with the simulation study conducted by Bernardo and Yu (2007) in that the number of underlying causal mutations, their effect sizes, and the heritability of the studied traits have a major influence on MAS and GS performance. In general, the consensus of these studies was that GS is better suited for polygenic traits, but, in practice, MAS could be used in conjunction with GS so that both large- and small-effect loci underlying trait variability could be captured. Although similar findings on the relative performance of MAS and GS are expected to be observed in *Miscanthus* × energycane hybrids, to the best of our knowledge, no studies have studied these two approaches in this system.

Given the inherent challenges with introgressing disease resistance from one perennial grass species to another, it is critical that the performance of GS and MAS is evaluated in the targeted species of our *Miscanthus* × energycane breeding program. Failure to do so could result in an inefficient breeding program that introduces an inadequate amount of disease resistance into energycane. Thus, the purpose of this study was to compare the ability of GS and MAS to predict traits simulated under different genetic architectures and marker densities. Due to the current lack of genomic data for *Miscanthus* × energycane hybrids, we conducted this analysis independently in *Miscanthus* × *Miscanthus* and sugarcane × sugarcane F₁ and BC₁ populations. Because these BC₁ populations are simulated, real phenotypic disease resistance data do not exist; thus, this study exclusively used simulated phenotypic data. We hypothesized that GS would outperform MAS for traits controlled by many genomic loci and that prediction accuracy would increase with marker density.

Materials and methods

Plant materials and genomic resources

This study evaluated the performance of GS and MAS in F₁ and subsequent simulated BC₁ populations in *Miscanthus* and sugarcane. The *Miscanthus* F₁ population was derived from a cross between *M. sinensis* ‘Kaskade’ and *M. sinensis* ‘Strictus’. *Miscanthus* is self-incompatible, and, consequently, the parents and F₁ progeny are highly heterozygous. Publicly available

genomic resources for this population include a set of Goldengate SNPs that were designed from variants discovered in RNA-seq data (Swaminathan et al. 2012), a set of SNPs that were discovered and called from RAD-seq data (Lu et al. 2013), and a composite genetic map (Liu et al. 2016). The markers used in this analysis comprised of 3044 RAD-Seq SNPs and 136 Goldengate SNPs (3180 SNPs in total) in 85 individuals. The genetic map was composed of 19 linkage groups.

The sugarcane F₁ population was composed of 173 individuals derived from a cross between sugarcane ‘CP95–1039’ × sugarcane ‘CP88–1762’ (Yang et al. 2017) and was genotyped with 21,895 single-dose SNP markers obtained using genotyping-by-sequencing (Yang et al. 2017; Elshire et al. 2011; Glaubitz et al. 2014). The resulting genomic data consisted of two parental genetic maps and their associated markers. The first parental genetic map (CP95–1039) consisted of 2453 markers spanning 162 LGs with a total map length of 4224.4 cm, while the second parental genetic map (CP88–1762) consisted of 2154 markers across 142 LGs with a total map length of 4373.2 cm. A composite genetic map was constructed using the LPmerge R package (Endelman and Plomion 2014), consisting of 162 linkage groups and 4607 markers (Yang et al. 2017).

Simulations

Using marker data from each F₁ population, we developed a simulation pipeline that allowed us to simulate phenotypes for each F₁ individual, compare the performance of GS to MAS, cross select F₁ individuals with a recurrent parent, and, finally, evaluate the performance of GS and MAS in the resulting BC₁ individuals. A schematic of this simulation pipeline is provided in Fig. 1, and critical relevant aspects are highlighted in the text below. The R scripts used for this simulation pipeline are available on GitHub at: <https://github.com/marcbios/Evaluation-of-Genomic-Selection-and-Marker-Assisted-Selection-in-Miscanthus-and-energycane>.

Phenotypic trait simulation

The same procedure was used to simulate traits in both F₁ and subsequent BC₁ populations. First, markers were randomly selected to be the quantitative trait nucleotides (QTNs), which constitute the genomic sources of

Fig. 1 Schematic depicting the simulation pipeline used to evaluate the performance of genomic selection (GS) and marker-assisted selection (MAS) in both F₁ and BC₁ populations. The gray boxes indicate the founders of the *M. sinensis* and sugarcane populations, i.e., ‘Kaskade’ and ‘Strictus’ for the *M. sinensis* populations, and ‘CP95-1039’ and ‘CP88-1762’ for the sugarcane

populations. The light blue box indicates the F₁ population, while the dark blue boxes depicts the BC₁ populations derived from GS and MAS. The red boxes indicate the evaluation of GS and MAS in the F₁ population using five-fold cross validation, and the black box shows the evaluation of the performance of GS and MAS in the BC₁ populations

variability for each trait. To assess the situation where causal mutations are not included in marker sets, QTNs in the *M. sinensis* populations were randomly selected from the 136 GoldenGate SNPs, while GS and MAS were conducted using the RAD-seq markers (3044 SNPs). Because only one marker set was available in the sugarcane populations, QTNs were randomly selected from all available 4607 markers.

The specific manner in which each randomly selected set of QTNs contributed to trait variability, which we defined as the genetic architecture of the trait, varied according the number of markers selected to be QTNs, the genetic mechanism of the QTN effects (i.e. additive, dominant, and epistatic), the size of the QTN effects, and heritability. A procedure similar to the ones described in Lande and Thompson (1990), Yu et al. (2008), and Bouchet et al. (2017) was used to simulate the effect sizes within each class of genetic effects so that they each followed a geometric series. Briefly, for a given class of p additive, dominant, or epistatic QTNs, the effect size was initiated as $\psi = \frac{p-1}{p+1}$, and then the

effect size of the i^{th} QTN was ψ^i . Thus the contribution of a set of p QTN to the phenotype was $\sum_{i=1}^p \psi^i x_{ij}$, where x_{ij} denotes the observed allele of the j^{th} individual at the i^{th} QTN, enumerated appropriately for quantifying additive, dominant, or epistatic effects. The numerical value of $\sum_{i=1}^p \psi^i x_{ij}$ constituted the baseline trait value for the j^{th} individual. The resulting genetic variance component of the trait σ_G^2 was calculated as the variance of the baseline trait. Finally, the non-genetic component of each trait was generated by simulating a normal random variable with mean 0 and variance $\sigma_\epsilon^2 = [\frac{\sigma_G^2}{H^2} - \sigma_G^2]$, where H^2 denotes the heritability. Such normal random variables were added to the baseline trait value for each individual. The specific values considered for the number of markers p selected to be QTNs, the genetic mechanism of the QTN effects, the size of the QTN effects, and heritability are summarized in Table 1. At each of the resulting 24 genetic architectures that were considered, a total of 100 replicate traits were simulated.

GS models

Genomic selection was performed using four GS models that were representative of the diversity of such models that have been explored in previous studies. These models were evaluated using the *sommer* (Covarrubias-Pazarán 2016), *BGLR* (Perez 2014), and *kernlab* (Karatzoglou et al. 2004) packages in the R programming language.

Additive, dominance, and epistasis models

The *sommer* R package was used to implement a GS model that simultaneously captures additive, dominance, and epistatic genomic signals. To achieve this, *sommer* fits mixed linear models that solve for multiple random effects with specific variance–covariance structures (Covarrubias-Pazarán 2016). The resulting “Additive+Dominance+Epistasis” (ADE) model we considered is written as follows:

$$Y = X\beta + u_A + u_D + u_E + \varepsilon, \quad (1)$$

where Y is a vector of observed phenotypic values; X is a design matrix; β is the vector of fixed effects; $u_A \sim \text{MVN}(0, A\sigma_A^2)$, $u_D \sim \text{MVN}(0, D\sigma_D^2)$, and $u_E \sim \text{MVN}(0, E\sigma_E^2)$ are vectors of individual random effects; A , D , and E are additive, dominance, and epistasis relationship matrices (VanRaden 2008; Su et al. 2012; Endelman 2013); and ε is the random error vector $\sim \text{MVN}(0, I\sigma_\varepsilon^2)$. The additive relationship matrix was estimated as follows:

$$A = \frac{Z_A Z_A'}{2 \sum_{j=1}^m p_j (1-p_j)} \quad (2)$$

Table 1 Description of the genetic architectures (genetic mechanism, heritability [H^2] and number of quantitative trait nucleotides [QTNs]) simulated in this study in both *M. sinensis* and sugarcane F_1 and BC_1 populations. The number in parentheses after each QTN number is the effect size of the largest QTN ψ , as described in the “Materials and methods”

QTN number	Genetic mechanism	H^2
4 (0.60), 10 (0.82), 20 (0.91), 100 (0.98)	Additive	0.7, 0.3
4 (0.60), 10 (0.82), 20 (0.91), 100 (0.98)	Dominance	0.7, 0.3
4 (0.60), 10 (0.82), 20 (0.91), 100 (0.98)	Epistatic	0.7, 0.3

where Z_A is an incidence matrix relating the additive contribution of each marker of each individual; the dominance relationship matrix was estimated as follows:

$$D = \frac{Z_D Z_D'}{2 \sum_{j=1}^m p_j q_j (1-p_j q_j)} \quad (3)$$

where Z_D is an incidence matrix relating the dominance contribution of each marker of each individual, and the epistasis relationship matrix was estimated as follows:

$$E_{aa} = A \# A \quad (\# \text{denotes the Hadamard product between two matrices}) \quad (4)$$

where p_j and q_j are the respective allelic frequencies for the j^{th} marker. These matrices were specified in *sommer* using the *A.mat*, *D.mat*, and *E.mat* functions.

BayesA

We used the *BGLR* package in R to implement BayesA. This model takes on the following form:

$$Y = 1\mu + \sum_{m=1}^s Z_m u_m + \varepsilon, \quad (5)$$

where Y is the vector of simulated phenotypic data; 1 is a vector of 1's; μ is the grand mean; s is the number of markers ($s > n$); Z_m is the observed genotypes at the m^{th} marker (coded as e.g. -1 , 0 , and 1 , where 1 and -1 were the homozygous classes and 0 the heterozygous class); u_m is the random effect of the m^{th} marker; and ε is the random error vector $\sim \text{MVN}(0, I\sigma_\varepsilon^2)$. The BayesA model assigns a scaled- t density prior to each of the random marker effects. All analyses conducted for this model used the default settings of 1000 burns and 2500 iterations.

RKHS

We also explored the reproducing kernel Hilbert space (RKHS) to determine if it yielded higher prediction accuracy in the presence of non-additive genomic signals. Specifically, Bayesian RKHS was analyzed in *BLGR*, which is described as follows:

$$Y = 1\mu + u + \varepsilon, \quad (6)$$

where Y is the vector of simulated phenotypic data; $\mathbf{1}$ is a vector of 1's; μ is the grand mean; u is vector of individual random effects $\sim \text{MVN}(0, K_h \sigma_u^2)$; and ε is the random error vector $\sim \text{MVN}(0, I \sigma_\varepsilon^2)$. For the Bayesian implementation of RKHS in *BLGR*, the priors $p(\mu, u, \varepsilon)$ are proportional to the product of the $\text{MVN}(0, K_h \sigma_u^2)$ and $\text{MVN}(0, I \sigma_\varepsilon^2)$ density functions. The matrix K_h is the kernel entries matrix with a Gaussian kernel, which uses the squared Euclidean distance between marker genotypes to quantify degree of relatedness between individuals, and a smoothing parameter h that multiplies each entry in K_h by a constant. Similar to the implementation of BayesA, the RKHS was conducted using the default 1000 burns, 2500 iterations, and the smoothing parameter h set to 0.5.

SVMR

Based on the support vector machine method developed by Vapnik (1995), support vector machine regression (SVMR) was initially applied to plant breeding by Maenhout et al. (2007). The main objective of SVMR is to minimize prediction error (Long et al. 2011). A detailed description of SVMR applied to GS is provided in Howard et al. (2014). In brief, SVMR is a kernel approach with a loss-function referred to as “ ε -intensive” meaning it is an approach that favors models that minimizes large residuals (Lorenz et al. 2011). In this analysis, we used the normal radial function kernel (*rbfdot*) implemented in the *ksvm* function of *kernelab* R package (Karatzoglou et al. 2004).

Five-fold cross validation

In this study, we used a five-fold cross validation approach to assess the ability of the tested GS models to accurately predict trait values in each of the F_1 populations. This approach has been previously described (Resende et al. 2012). In brief, both genomic and phenotypic information of all the individuals were divided into five approximately equally-sized groups. A set of four groups were used as a training set to fit the GS model, while the remaining one group was used as the evaluation or test set. Prediction accuracy was quantified as the Pearson correlation between the simulated trait values and the genomic estimated breeding values (GEBVs) predicted from a given GS model evaluated in the test set. The process was repeated five times to

ensure that each of the five groups was used as a test set. All four GS models were evaluated in the same folds to facilitate comparison.

Further evaluation of GS in the F_1 populations

Several other metrics besides the average Pearson's correlation coefficient across folds were used to assess the performance of four GS models in the two F_1 populations. To evaluate the impact of marker density on GS performance, all analyses were conducted using the full complement of markers, as well as on random samples of 0.25, 0.50, and 0.75 of markers. In addition, a regression model was fitted with the observed values as the response variable and the GEBVs as the explanatory variable to assess the bias of the GEBVs, as described in Daetwyler et al. (2013). Finally, we calculated the coincidence index (Hamblin and Zimmermann 2011; Fernandes et al. 2018) to evaluate the proportion of individuals that overlap between individuals with highest trait values (10%) in simulated phenotypes and predicted phenotypic trait values for the validation population.

MAS in the F_1 populations

We carried out MAS in the two F_1 populations (*M. sinensis* and sugarcane) by using the unified mixed linear model (MLM; Yu et al. 2006) in the genome association and prediction integrated tool (GAPIT; Lipka et al. 2012) R package to perform a genome-wide association study (GWAS) within each training set generated from the five-fold cross validation approach. Here, the same folds used to evaluate the performance of GS were used. The top four GWAS signals with the strongest marker-trait associations from each training set were selected and used to obtain GEBVs of F_1 individuals in the test set, and subsequently perform MAS. The same previously described approaches used to evaluate prediction accuracy, bias, and coincidence were then implemented to evaluate the performance of MAS in these F_1 populations.

Simulation of BC_1 populations

At this stage of the simulation pipeline, we backcrossed all F_1 individuals with the top 10% highest GEBVs from each of the four GS models and MAS. Specifically, each of these selected F_1 individuals from the *M. sinensis*

‘Kaskade’ $\times$ *M. sinensis* ‘Strictus’ F_1 population was backcrossed to *M. sinensis* ‘Kaskade’ to create *M. sinensis* ‘Kaskade’- BC_1 individuals. Likewise, each of the selected F_1 individuals from the sugarcane ‘CP95–1039’ $\times$ sugarcane ‘CP88–1762’ F_1 population was backcrossed to sugarcane ‘CP95–1039’ to create sugarcane ‘CP95–1039’- BC_1 individuals. Simulations were based on the assumption of normal pairing of homologous chromosomes (normal meiotic segregation) in *M. sinensis* and sugarcane since the genomic data were made up of mainly single dose markers (Mollinari and Garcia 2018). To simulate new genomes for each BC_1 individual, we first used a custom Haldane’s mapping function in R to simulate crossover events along the genomes of the selected F_1 individuals and recurrent parent. Next, we subdivided each chromosome of each selected F_1 individual and the recurrent parent into two gametes. One gamete was then randomly selected from each parent (i.e., a selected F_1 individual and the recurrent parent) to form the genome of the corresponding BC_1 progeny. Thus, the genome of this BC_1 individual had genomic information contributed by both the recurrent parent and the selected F_1 individual. Using this approach, we ended up with four GS derived BC_1 populations (i.e. one BC_1 population for each GS model) and one MAS derived BC_1 population for each genetic architecture simulated in each species.

Phenotypic evaluation of GS and MAS in the BC_1 population

In each BC_1 population, we simulated traits with the exact same genetic architecture (i.e., with the same markers selected to be QTNs) used for simulating the F_1 populations (summarized in Table 1). To assess the capability of an entire F_1 population to predict breeding values in each GS-derived BC_1 population, the corresponding GS model was fitted to one of the 100 replicate traits simulated at each genetic architecture in the F_1 population. Similarly, a GWAS for the same replicate traits was conducted in the F_1 population using the unified MLM (Yu et al. 2006). We then conducted MAS in the MAS-derived BC_1 population using the top four peak-associated GWAS markers. Prediction accuracy was estimated as the Pearson correlation between the trait values simulated in the BC_1 individuals and the corresponding GEBVs.

Results

In general, consistent results were obtained between high and low heritability traits, although the low heritability traits yielded lower prediction accuracies. Therefore, for brevity, the results for high heritability traits are presented below unless otherwise noted.

Effect of marker densities on GS in F_1 populations (*M. sinensis* and sugarcane)

To investigate the effect of marker density on genomic prediction, different proportions of genome-wide markers were used for prediction in both *M. sinensis* and sugarcane F_1 populations. In *M. sinensis*, the only genetic architectures where a discernable impact of marker density was observed on prediction accuracies of several GS models was for traits with either 100 additive or 100 epistatic QTN (Fig. 2). For these genetic architectures, the prediction accuracies of all models except for SVR tended to increase as the marker density increased. In contrast, there were no differences in prediction accuracies across marker densities for all traits simulated with four QTNs under high heritability in the additive mechanism (Fig. 2a). For traits simulated with four epistatic QTNs, the impact of marker densities on prediction accuracies varied among GS models (Fig. 2b). For instance, in the ADE, RKHS, and SVR models, prediction accuracies increased with greater marker density but decreased for the Bayes A GS model (Fig. 2b). In both the additive and epistasis mechanisms, there were no differences in prediction accuracies among marker densities for traits simulated with 4 and 100 QTNs under low heritability (Online Resource 1). In sugarcane, a discernable pattern was only observed in traits simulated with four additive QTN under high heritability using the ADE, RKHS, and SVR GS models. Here prediction accuracy increased as marker density increased from 0.25 to 0.75 of the genetic markers, while prediction accuracy decreased to approximately the same accuracy that was observed with 0.50 of the genetic markers when all genetic markers were considered (Online Resource 1).

Comparison of GS and MAS models in the *M. sinensis* F_1 population

For each cross-validation fold across the 100 replicates of each trait simulated in the *M. sinensis* F_1

Fig. 2 Effect of marker density of prediction accuracy for traits simulated with additive (**a**) and epistatic (**b**) quantitative trait nucleotides (QTNs) in the *M. sinensis* F₁ population. **a** Results for traits simulated with additive QTNs. Box plots showing the distribution of five-fold cross validation prediction accuracies (Y-axis, estimated as the correlation between simulated trait values and predicted trait values across five folds) for different marker densities evaluated using three four GS models (X-axis) [Additive-

Dominance-Epistasis Kernel Model (ADE), BayesA (BA), and reproducing kernel Hilbert space (RKHS), and support vector regression (SVR)]. Here, 100 replicate traits with high heritability (0.7) were simulated. The left graph shows the results for traits simulated with four additive QTN, while the right graphs shows the results for traits simulated with 100 additive QTN. **b** Results for traits simulated with epistatic QTNs, presented in a manner that is identical to the results presented in part (**a**)

population, we compared the prediction accuracy between the four GS models and MAS. In general, MAS yielded equivalent or higher prediction accuracies compared with the GS models for the smallest number of simulated QTN, but, as the number QTN underlying the simulated traits increased, the

prediction accuracy of MAS decreased relative to the GS models (Fig. 3a and Online Resource 2: Tables S1 and S2 showing the results for the corresponding traits simulated with additive, dominance, and epistatic QTNs). We next evaluated the bias and coincidence index of each GS and MAS model. For

all high heritability traits, ADE, BayesA, and MAS had the least bias. Moreover, we observed that prediction accuracies were less biased under high heritability compared to low heritability (Online Resource 2: Tables S3, S4, S5, and S6). Interestingly, the ADE GS models infrequently gave extremely biased GEBVs, especially for low-heritability traits (Online Resource 2: Tables S4 and S6).

The analysis of coincidence indices indicated that the ability of the GS and MAS models to identify the individuals with optimal simulated trait values (i.e., observed trait values generated from the simulations) depended mostly upon the number of simulated QTN (Online Resource 2: Tables S7 and S8). However, for high heritability genetic architectures the ADE and BayesA GS models tended to consistently have higher coincidence indices, indicating that the individuals with the top 10% GEBVs from these two approaches were the most consistent with individuals in the top 10% of simulated trait values. Additionally, we observed that for traits with 100 QTN, MAS had noticeably lower coincidence indices than the GS models.

Comparison between GS and MAS in the sugarcane F_1 population

In the sugarcane F_1 population, prediction accuracies across 100 replicates were compared between the tested GS and MAS approaches. Similar to the results in the *M. sinensis* F_1 population, the performance of MAS tended to get worse as the number of simulated QTN increased (Fig. 3b, Online Resource 2: Tables S9 and S10). Unlike the results from the similar analysis conducted in the *M. sinensis* F_1 population, BayesA consistently yielded among the highest prediction accuracies compared to other GS models and MAS, even for the genetic architectures with four QTN. Moreover, the phenomenon of the ADE GS model infrequently producing extremely biased GEBVs in the *M. sinensis* F_1 population (Online Resource 2: Table S3–S6) was notably absent in the sugarcane F_1 population (Online Resource 2: Table S11–S14). The analysis of coincidence indices showed that MAS tended to select individuals that were the least consistent with those that had the top 10% simulated trait values, while BayesA tended to select individuals that were the most consistent (Online Resource 2: Table S15–S16).

Comparison between MAS and GS models in MAS and GS derived BC_1 populations

Within each evaluated species, five backcross populations (ADE BC_1 , BayesA BC_1 , RKHS BC_1 , SVR BC_1 and MAS BC_1) were generated by crossing genotypes selected using each GS model and MAS in the F_1 to the recurrent parent. In general, these results suggest that GEBVs obtained from either training a GS model or identifying peak-associated markers in an F_1 population can accurately predict values of traits controlled by additive, epistatic, and a small number of dominance QTN in BC_1 progeny (Fig. 4). However, the results also suggest that such an approach cannot accurately predict trait values of BC_1 individuals when the underlying genetic architecture consists of a large number of dominance QTN (Fig. 4). For both species, no discernable pattern was observed among the performance of GS models in the BC_1 populations. Interestingly, contrasting results on the performance of MAS were obtained between the two species. In the *M. sinensis* BC_1 populations, MAS typically yielded among the lowest prediction accuracies across the tested approaches (Fig. 4a, Online Resource 2: Table S17–S18). In contrast the results in the sugarcane BC_1 populations suggest that MAS may be well-suited for genetic architectures consisting of a relatively small number of additive QTNs (Fig. 4b, Online Resource 2: Table S19–S20).

Discussion

To ascertain the potential of GS and MAS to facilitate introgression of disease resistance into energycane, we used marker data from F_1 populations in *M. sinensis* and sugarcane to simulate traits under a wide variety of genetic architectures. Although there were no GS models that consistently outperformed others, we observed that GS tended to yield higher prediction accuracies, select more individuals with the peak-performing simulated trait values, and more accurately predict trait values of BC_1 progeny relative to MAS. Coupled with the finding that at best the prediction accuracy of MAS declined as the number of simulated QTNs underlying a trait increased, our results suggest that GS is better suited for introgressing genetic sources of horizontal disease resistance into energycane. That is, our study demonstrates that when many loci underlie the genetic sources of variation, accounting for both large and small

Fig. 3 Comparison of prediction accuracies between genomic selection (GS) and marker-assisted selection (MAS) for traits simulated with four, 10, 20, and 100 QTNs in the additive mechanism under high heritability (0.7) in the *M. sinensis* × *M. sinensis* (a) and sugarcane × sugarcane (b) F₁ populations. **a** Prediction accuracies for the *M. sinensis* × *M. sinensis* F₁ population. The GS models [Additive-Dominance-Epistasis Kernel Model (ADE), Bayes A (BA), reproducing kernel Hilbert space (RKHS), and support vector regression (SVR)] and marker-assisted selection (MAS) are presented on the X-axis. The prediction accuracy, estimated as the correlation between simulated trait values and predicted trait values across five folds and 100 replications, is presented on the Y-axis. The title of each graph indicated the number of quantitative trait nucleotides (QTNs) underlying the simulated traits. **b** Prediction accuracies for the sugarcane × sugarcane F₁ population, presented in a manner that is identical to the results presented in part (a)

effect loci are critical for maximizing prediction accuracy and selecting optimal F₁ individuals to backcross. However, when disease resistance is controlled by only a small number of genes, our results suggest that MAS performs adequately.

Impact of data characteristics on GS and MAS performance

One important characteristic that sets the data we analyzed apart from those assessed in prior work in other species is the small number of individuals ($n = 85$ in the *M. sinensis* F₁ population and $n = 173$ in the sugarcane

Fig. 4 Comparison of observed simulated traits with high heritability (0.7) and corresponding genomic estimated breeding values (GEBVs) in the BC₁ populations generated in the *M. sinensis* × *M. sinensis* (a) and sugarcane × sugarcane (b) Population. **a** Results for the BC₁ populations generated from the *M. sinensis* × *M. sinensis* populations. The genetic architectures of the traits presented are subdivided by the genetic mechanisms (rows) and number of quantitative trait nucleotides (QTNs; columns). The GS models [Additive-Dominance-Epistasis Kernel Model (ADE), Bayes A (BA), reproducing kernel Hilbert space (RKHS), and support vector regression (SVR)] and marker-assisted selection

(MAS) are presented on the X-axis. The trait values are presented on the Y-axis. The pink and blue box plots respectively show the distribution of observed simulated trait values and the corresponding GEBVs, which were obtained by fitting the corresponding GS models (for ADE, BA, RKHS and SVR) or conducting a genome-wide association study in the F₁ population. The numbers above each pair of box plots indicate the Pearson correlation coefficient between the observed simulated trait values and the GEBVs. **b** Results for the BC₁ populations generated from the sugarcane × sugarcane populations presented in a manner that is identical to the results presented in part (a)

F₁ population). Given that these sizes are substantially smaller than data sets evaluated in other studies (Spindel et al. 2015; Battenfield et al. 2016; Fernandes et al. 2018), the data we analyzed provided a unique opportunity to explore the effectiveness of GS and MAS under such constraints. Such constraints are likely to be more realistic of the introgression of genes from wide crosses, where interspecific incompatibilities can result in low numbers of early generation progeny. In this regard, the ability to demonstrate the superior performance of GS over MAS was illuminating. Undoubtedly, the feasibility of obtaining high-throughput genotype and phenotype data afforded by advances in next-generation sequencing, unmanned aerial systems, and other high-throughput phenotyping platforms (Huang et al. 2009; Rife et al. 2011; Elshire et al. 2011; Haghighattalab et al. 2016; Wang et al. 2018) will result in larger quantities of data becoming available in energycane and *M. sinensis*, and our expectations would be that GS will show the same advantages over MAS that have been demonstrated in this work and in other species. Nevertheless, these results aid current breeding efforts because we show that GS appears to be superior to MAS, even when small data sets are used.

Given that two marker sets were available in the *M. sinensis* F₁ population, we were able to simulate QTN using one marker set and evaluate the performance of GS and MAS in the other. This is beneficial because it likely reflects the reality that many causal mutations underlying traits are not included in marker data sets, and these markers are typically not in complete linkage disequilibrium with them (Korte and Farlow 2013). Coupled with the markers in the *M. sinensis* population used to evaluate GS and MAS being generated from RAD-seq technology, this assessment is directly applicable to our anticipated follow-up studies where we will strictly use sequence-based markers.

One drawback of this work is that only simulated phenotypic data were used. No actual disease resistance traits were considered because no such data exist for all of the populations we evaluated. That is, the BC₁ populations we evaluated were simulated; no actual BC₁ populations generated in this manner exist in real life. Thus, it is currently impossible to confirm the results of our simulation study using real disease resistance traits. This lack of data was one of the major factors underlying our decision to simulate traits with a wide variety of contrasting genetic architectures, thereby providing guidance in the absence of empirical

results from these populations and new hypotheses to test in future experiments.

Performance of GS versus MAS

Our results demonstrated that the prediction accuracy of MAS tended to decline as the number of simulated QTN increased. This finding is similar to the simulation conducted in Bernardo and Yu (2007), which showed that the response to GS was higher than MAS for traits controlled by 20–100 loci. Moreover, previous studies using real phenotypic crop data agree with these results: GS outperformed MAS for horizontal disease resistance in wheat (Arruda et al. 2015) and polygenic nutrient quantity traits in rye (Wang et al. 2014), while Owens et al. (2014) showed that GS and MAS could perform equally well for the relatively simpler carotenoid traits in maize. Combined with our finding that the GS models tended to give higher coincidence indices than the MAS models, we conclude that, while MAS may perform adequately for certain genetic architectures, the tested GS models should consistently perform optimally across a wider range of traits. Nonetheless, for breeding programs focused specifically on traits controlled by very few, well-defined loci with large effect sizes (such as some disease resistance traits), the lower cost and technical overhead of MAS, combined with similar performance, is likely to remain the preferred option. This recommendation is consistent with the findings of Olatoye et al. (2019), which used real cowpea phenotypic data to show that GS offered no advantages in prediction accuracy over MAS for traits where the strongest associated marker explained between 10 and 21% of the total phenotypic variation.

Across all of the simulation settings that were explored, none of the GS models decisively outperformed the others. That is, for a given genetic architecture, the prediction accuracies, bias, and coincidence indices of the four studied GS models were generally similar to each other. In general, this result is consistent with previous studies conducted in other species (e.g., Heslot et al. 2012). However, we noticed several genetic architectures where certain GS models were either the best- or worst-performing models. For example, our finding that the ADE model occasionally gave extremely biased GEBVs in the *M. sinensis* F₁ population for low-heritable simulated traits may suggest that this model should not be used in practice for such traits. On the other hand, our results show that BayesA often yielded

the highest prediction accuracies (especially in the sugarcane F_1 population), the highest coincidence indices, and the lowest biases. Therefore, our recommendation is to run several GS models and use the collective results to identify individuals in which to backcross. Given that all four of the GS models we explored can be fitted to data and tested using publicly available software, undertaking such a multifaceted analysis should be feasible.

To summarize, the main findings from these simulation studies on the performance of GS and MAS in *M. sinensis* and sugarcane are consistent with similar work conducted in other species. Within the context of our specific *M. sinensis* $\times$ energycane breeding program, our simulation studies help underscore that either GS, MAS, or both should be successful in expediting the introgression of disease resistance from *M. sinensis* into energycane. Because we simulated traits with various sizes of additive, dominance, and two-way epistatic QTN across two different heritabilities, we expect these studies to foreshadow the performance of MAS and GS in our breeding material, particularly with respect to which genetic architectures they are expected to perform optimally. In these regards, the findings from the simulation studies presented here are crucial for the future success of this *M. sinensis* $\times$ energycane breeding program.

Conclusions and recommendations

Marker-assisted breeding efforts for introgressing disease resistance into energycane need to use statistical models that yield the highest prediction accuracies for both horizontal and vertical disease resistance. In this regard, we used genomic marker data currently available in both sugarcane and *M. sinensis* to demonstrate that the performance of GS is superior to MAS in the great majority of simulated genetic architectures. That is, while MAS performed reasonably well for traits controlled by a smaller number of QTN, the prediction accuracy of MAS dropped dramatically compared with the four tested GS models as the number of underlying QTN increased. We therefore conclude that MAS is a reasonable option for advancing vertical disease resistance. In contrast, we recommend that GS be used at the forefront of breeding efforts for horizontal disease resistance in energycane. As more attention and resources for genotyping, phenotyping, and sampling more individuals gets directed towards energycane and related species like *M. sinensis*, we hypothesize that the increase in

number of markers and precision in quantifying traits will reveal novel genomic regions with intricate contributions to trait variability. Thus, we would anticipate that observed differences between the performance of GS models and MAS to increase, as marker and phenotypic data for *M. sinensis* $\times$ energycane populations becomes more affordable and extensive.

Acknowledgments This study is made possible by the support of DOE Office of Science, Office of Biological and Environmental Research (BER), grant no. DE-SC0016264. The contents of this manuscript are the sole responsibility of the authors. We acknowledge the efforts of Dr. Matt Hudson, Dr. Samuel Fernandes, and Brian Rice for reading through the manuscript and providing critical feedback.

Compliance with ethical standards

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Annicchiarico P, Nazzicari N, Li X, Wei Y, Pecetti L, Brummer EC (2015) Accuracy of genomic selection for alfalfa biomass yield in different reference populations. *BMC Genomics* 16: 1020. <https://doi.org/10.1186/s12864-015-2212-y>
- Arruda MP, Brown PJ, Lipka AE et al (2015) Genomic selection for predicting head blight resistance in a wheat breeding program. *Plant Genome* 8:1–12. <https://doi.org/10.3835/plantgenome2015.01.0003>
- Battenfield SD, Guzmán C, Gaynor RC et al (2016) Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. *Plant Genome* 9:0. <https://doi.org/10.3835/plantgenome2016.01.0005>
- Beavis WD (1998) QTL analyses: power, precision, and accuracy. In: *Molecular dissection of complex traits*. CRC Press, New York

- Beavis WD (1995) The power and deceit of QTL experiments: lessons from comparative QTL studies. In: Proceedings of the Forty-ninth Annual Corn and Sorghum Industry Research Conference. ASTA, Washington, pp 252–268
- Ben-Ari G, Lavi U (2012) Marker-assisted selection in plant breeding. In: Plant biotechnology and agriculture. Elsevier, pp 163–184
- Bernardo R (2001) What if we knew all the genes for a quantitative trait in hybrid crops? *Crop Sci* 41:1–4. <https://doi.org/10.2135/cropsci2001.4111>
- Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. *Crop Sci* 47:1082. <https://doi.org/10.2135/cropsci2006.11.0690>
- Beyene Y, Semagn K, Mugo S et al (2015) Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Sci* 55:154–163. <https://doi.org/10.2135/cropsci2014.07.0460>
- Bouchet S, Olatoye MO, Marla SR et al (2017) Increased power to dissect adaptive traits in global sorghum diversity using a nested association mapping population. *Genetics* 206:573–585. <https://doi.org/10.1534/genetics.116.198499>
- Cao SL, Loladze A, Yuan YB et al (2017) Genome-wide analysis of tar spot complex resistance in maize using genotyping-by-sequencing SNPs and whole-genome prediction. *Plant Genome* 10. <https://doi.org/10.3835/plantgenome2016.10.0099>
- Cerrudo D, Cao S, Yuan Y, Martinez C, Suarez EA, Babu R, Zhang X, Trachsel S (2018) Genomic selection outperforms marker assisted selection for grain yield and physiological traits in a maize doubled haploid population across water treatments. *Front Plant Sci* 9:366. <https://doi.org/10.3389/fpls.2018.00366>
- Chen YH, Lo CC (1989) Disease resistance and sugar content in *Saccharum-Miscanthus* hybrids. *Rep Taiwan Sugar Res Inst* 36:1–7
- Costet L, Le Cunff L, Royaert S et al (2012) Haplotype structure around Br1 reveals a narrow genetic basis for brown rust resistance in modern sugarcane cultivars. *Theor Appl Genet* 125:825–836. <https://doi.org/10.1007/s00122-012-1875-x>
- Covarrubias-Pazarán G (2016) Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One* 11:1–15. <https://doi.org/10.1371/journal.pone.0156744>
- Daetwyler HD, Calus MP, Pong-Wong Ricardo, de los Campos G, and Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193(2):347–365
- de Almeida Costa PM (2015) Prediction of breeding values in sugarcane using pedigree and genomic information. Universidade Federal de Viçosa
- Dong H, Liu S, Clark LV et al (2018) Genetic mapping of biomass yield in three interconnected *Miscanthus* populations. *GCB Bioenergy* 10:165–185. <https://doi.org/10.1111/gcbb.12472>
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. <https://doi.org/10.1371/journal.pone.0019379>
- Endelman J (2013) Genomic prediction with rrBLUP 4 Jeffrey Endelman June 15, 2013. 1–9
- Endelman JB, Plomion C (2014) LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics* 30:1623–1624. <https://doi.org/10.1093/bioinformatics/btu091>
- Fernandes SB, Dias KOG, Ferreira DF, Brown PJ (2018) Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theor Appl Genet* 131:747–755. <https://doi.org/10.1007/s00122-017-3033-y>
- Gezan SA, Osorio LF, Verma S, Whitaker VM (2017) An experimental validation of genomic selection in octoploid strawberry. *Hortic Res* 4:16070. <https://doi.org/10.1038/hortres.2016.70>
- Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9:e90346. <https://doi.org/10.1371/journal.pone.0090346>
- 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. <https://doi.org/10.1007/s00122-013-2156-z>
- Haghighattalab A, González Pérez L, Mondal S, Singh D, Schinstock D, Rutkoski J, Ortiz-Monasterio I, Singh RP, Goodin D, Poland J (2016) Application of unmanned aerial systems for high throughput phenotyping of large wheat breeding nurseries. *Plant Methods* 12:35. <https://doi.org/10.1186/s13007-016-0134-6>
- Hamblin J, de Oliveira Zimmermann MJ (2011) Breeding common bean for yield in mixtures. In: *Plant Breeding Reviews*. John Wiley & Sons, Inc., Hoboken, pp 245–272
- Heffner EL, Jannink J-L, Iwata H et al (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci* 51:2597. <https://doi.org/10.2135/cropsci2011.05.0253>
- Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. *Crop Sci* 52:146–160. <https://doi.org/10.2135/cropsci2011.06.0297>
- Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and Epistatic genetic architectures. *G3* 4:1027–1046. <https://doi.org/10.1534/g3.114.010298>
- Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T, Dong G, Sang T, Han B (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res* 19:1068–1076. <https://doi.org/10.1101/gr.089516.108>
- Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. *Briefings Funct Genomics Proteomics* 9:166–177. <https://doi.org/10.1093/bfpg/elq001>
- Jiang G-L (2013) Molecular markers and marker-assisted breeding in plants. In: *Plant breeding from laboratories to fields* InTech
- Kandel R, Yang X, Song J, Wang J (2018) Potentials, challenges, and genetic and genomic resources for sugarcane biomass improvement. *Front Plant Sci* 9:151. <https://doi.org/10.3389/fpls.2018.00151>
- Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) Kernlab - an S4 package for kernel methods in R. *J Stat Softw* 11:1–20. <https://doi.org/10.18637/jss.v011.i09>

- Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9:29. <https://doi.org/10.1186/1746-4811-9-29>
- Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–756. <https://doi.org/10.1046/j.1365-2540.1998.00308.x>
- Lipka AE, Tian F, Wang Q et al (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28:2397–2399. <https://doi.org/10.1093/bioinformatics/bts444>
- Liu S, Clark LV, Swaminathan K et al (2016) High-density genetic map of *Miscanthus sinensis* reveals inheritance of zebra stripe. *GCB Bioenergy* 8:616–630. <https://doi.org/10.1111/gcbb.12275>
- Long N, Gianola D, Rosa GJM, Weigel KA (2011) Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor Appl Genet* 123:1065–1074. <https://doi.org/10.1007/s00122-011-1648-y>
- Lorenz AJ, Chao S, Asoro FG et al (2011) Chapter 2: genomic selection in plant breeding: knowledge and prospects. *Adv Agron* 110:77–123. <https://doi.org/10.1016/B978-0-12-385531-2.00002-5>
- Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, Buckler ES, Costich DE (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet* 9:e1003215. <https://doi.org/10.1371/journal.pgen.1003215>
- Maenhout S, De Baets B, Haesaert G, Van Bockstaele E (2007) Support vector machine regression for the prediction of maize hybrid performance. *Theor Appl Genet* 115:1003–1013. <https://doi.org/10.1007/s00122-007-0627-9>
- Massman JM, Jung HJG, Bernardo R (2013) Genomewide selection versus marker-assisted recurrent selection to improve grain yield and Stover-quality traits for cellulosic ethanol in maize. *Crop Sci* 53:58–66. <https://doi.org/10.2135/cropsci2012.02.0112>
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829. Doi: 11290733
- Mirdita V, He S, Zhao Y, Korzun V, Bothe R, Ebmeyer E, Reif JC, Jiang Y (2015) Potential and limits of whole genome prediction of resistance to *Fusarium* head blight and *Septoria tritici* blotch in a vast central European elite winter wheat population. *Theor Appl Genet* 128:2471–2481. <https://doi.org/10.1007/s00122-015-2602-1>
- Mollinari M, Garcia AAF (2018) Linkage analysis and haplotype phasing in experimental autopolyploid populations with high ploidy level using hidden Markov models *bioRxiv*:415232. <https://doi.org/10.1101/415232>
- Okogbenin E, Porto MCM, Egesi C et al (2007) Marker-assisted introgression of resistance to cassava mosaic disease into latin American germplasm for the genetic improvement of cassava in Africa. *Crop Sci* 47:1895–1904. <https://doi.org/10.2135/cropsci2006.10.0688>
- Olatoye MO, Hu Z, Aikpokpodion PO (2019) Epistasis detection and modeling for genomic selection in cowpea (*Vigna unguiculata* L. Walp.). *Front Genet* 10:677
- Owens BF, Gore MA, Magallanes-Lundback M et al (2014) A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. *Genetics* 198:1699–1716. <https://doi.org/10.1534/genetics.114.169979>
- Perez P (2014) BGLR : a statistical package for whole genome regression and prediction. *Genetics* 198:483–495. <https://doi.org/10.1534/genetics.114.164442>
- Poland J (2015) Breeding-assisted genomics. *Curr Opin Plant Biol* 24:119–124. <https://doi.org/10.1016/j.pbi.2015.02.009>
- Resende JFR, Muñoz P, Resende MDV et al (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). *Genetics* 190:1503–1510. <https://doi.org/10.1534/genetics.111.137026>
- Rife TW, Wu S, Bowden RL, Poland JA (2011) Spiked GBS: a unified, open platform for single marker genotyping and whole-genome profiling. <https://doi.org/10.1186/s12864-015-1404-9>
- Rutkoski J, Benson J, Jia Y et al (2012) Evaluation of genomic prediction methods for *Fusarium* head blight resistance in wheat. *Plant Genome J* 5:51. <https://doi.org/10.3835/plantgenome2012.02.0001>
- Rutkoski J, Singh RP, Huerta-Espino J et al (2015) Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. *Plant Genome* 8:0. <https://doi.org/10.3835/plantgenome2014.10.0074>
- Saha MC, Bhandhari HS, Bouton JH (ed) (2013) *Bioenergy Feedstocks: Breeding and Genetics*. John Wiley & Sons, Danvers, MA
- Spindel J, Begum H, Akdemir D et al (2015) Genomic selection and association mapping in Rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of Rice genomic selection in elite, tropical Rice breeding lines. *PLoS Genet* 11:1–25. <https://doi.org/10.1371/journal.pgen.1004982>
- Su G, Christensen OF, Ostensen T et al (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PLoS One* 7:e45293. <https://doi.org/10.1371/journal.pone.0045293>
- Swaminathan K, Chae W, Mitros T, Varala K, Xie L, Barling A, Glowacka K, Hall M, Jezowski S, Ming R, Hudson M, Juvik JA, Rokhsar DS, Moose SP (2012) A framework genetic map for *Miscanthus sinensis* from RNAseq-based markers shows recent tetraploidy. *BMC Genomics* 13:142. <https://doi.org/10.1186/1471-2164-13-142>
- Technow F, Bürger A, Melchinger AE (2013) Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3 (Bethesda)* 3:197–203. <https://doi.org/10.1534/g3.112.004630>
- Tew TL, Cobill RM (2008) *Genetic improvement of sugarcane (Saccharum spp.) as an energy crop*. Springer Science & Business Media, New York
- Utz HF, Melchinger AE, Schön CC (2000) Bias and sampling error of the estimated proportion of genotypic variance explained by quantitative trait loci determined from experimental data in maize using cross validation and validation with independent samples. *Genetics* 154:1839–1849. <https://doi.org/10.2307/1403680>
- VanRaden PM (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>

- Vapnik VN (1995) The nature of statistical learning theory, 1st edn. Springer, New York
- Wang X, Li W, Huang Y et al (2013) Evaluation of sugarcane introgression lines for resistance to brown rust disease caused by *Puccinia melanocephala*. *Trop Plant Pathol* 38:97–101. <https://doi.org/10.1590/S1982-56762013000200002>
- Wang X, Singh D, Marla S, Morris G, Poland J (2018) Field-based high-throughput phenotyping of plant height in sorghum using different sensing technologies. *Plant Methods* 14:53. <https://doi.org/10.1186/s13007-018-0324-5>
- Wang Y, Mette MF, Miedaner T et al (2014) The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years *BMC Genomics* 15. <https://doi.org/10.1186/1471-2164-15-556>
- Xu S (2003) Theoretical basis of the Beavis effect. *Genetics* 165: 2259–2268
- Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci* 48:391. <https://doi.org/10.2135/cropsci2007.04.0191>
- Yang X, Islam MS, Sood S et al (2018) Identifying quantitative trait loci (QTLs) and developing diagnostic markers linked to Orange rust resistance in sugarcane (*Saccharum* spp.). *Front plant Sci* 9:350. <https://doi.org/10.3389/fpls.2018.00350>
- Yang X, Sood S, Glynn N et al (2017) Constructing high-density genetic maps for polyploid sugarcane (*Saccharum* spp.) and identifying quantitative trait loci controlling brown rust resistance. *Mol breed* 37:116. <https://doi.org/10.1007/s11032-017-0716-7>
- Yohannes T, Tesfamichael A, Kiambi D et al (2015) Marker-assisted introgression improves *Striga* resistance in an Eritrean farmer-preferred *Sorghum* variety. *F Crop Res* 173: 22–29. <https://doi.org/10.1016/j.fcr.2014.12.008>
- Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178:539–551. <https://doi.org/10.1534/genetics.107.074245>
- Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen M, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208. <https://doi.org/10.1038/ng1702>
- Zhang X, Pérez-Rodríguez P, Semagn K, Beyene Y, Babu R, López-Cruz MA, San Vicente F, Olsen M, Buckler E, Jannink JL, Prasanna BM, Crossa J (2015) Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity (Edinb)* 114:291–299. <https://doi.org/10.1038/hdy.2014.99>
- Zhao Y, Mette MF, Gowda M et al (2014) Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity (Edinb)* 112: 638–645. <https://doi.org/10.1038/hdy.2014.1>

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