

# Genomic prediction applied to high-biomass sorghum for bioenergy production

Amanda Avelar de Oliveira · Maria Marta Pastina · Vander Filipe de Souza ·  
Rafael Augusto da Costa Parrella · Roberto Willians Noda ·  
Maria Lúcia Ferreira Simeone · Robert Eugene Schaffert ·  
Jurandir Vieira de Magalhães · Cynthia Maria Borges Damasceno ·  
Gabriel Rodrigues Alves Margarido 

Received: 10 August 2017 / Accepted: 13 March 2018 / Published online: 10 April 2018  
© The Author(s) 2018

**Abstract** The increasing cost of energy and finite oil and gas reserves have created a need to develop alternative fuels from renewable sources. Due to its abiotic stress tolerance and annual cultivation, high-biomass sorghum (*Sorghum bicolor* L. Moench) shows potential as a bioenergy crop. Genomic selection is a useful tool for accelerating genetic gains and could restructure plant breeding programs by enabling early selection and reducing breeding cycle duration. This work aimed at predicting breeding values via genomic selection models for 200 sorghum genotypes comprising landrace accessions and breeding lines from biomass and saccharine groups. These genotypes were divided into two sub-panels, according to breeding purpose. We evaluated the following phenotypic biomass traits: days to flowering, plant height, fresh and dry matter yield, and fiber, cellulose, hemicellulose, and lignin proportions. Genotyping by sequencing yielded more than

258,000 single-nucleotide polymorphism markers, which revealed population structure between sub-panels. We then fitted and compared genomic selection models BayesA, BayesB, BayesC $\pi$ , BayesLasso, Bayes Ridge Regression and random regression best linear unbiased predictor. The resulting predictive abilities varied little between the different models, but substantially between traits. Different scenarios of prediction showed the potential of using genomic selection results between sub-panels and years, although the genotype by environment interaction negatively affected accuracies. Functional enrichment analyses performed with the marker-predicted effects suggested several interesting associations, with potential for revealing biological processes relevant to the studied quantitative traits. This work shows that genomic selection can be successfully applied in biomass sorghum breeding programs.

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s11032-018-0802-5>) contains supplementary material, which is available to authorized users.

A. A. de Oliveira · G. R. A. Margarido (✉)  
Department of Genetics, Luiz de Queiroz College of Agriculture,  
University of São Paulo, Piracicaba, SP 13418-900, Brazil  
e-mail: gramarga@usp.br

M. M. Pastina · V. F. de Souza · R. A. da Costa Parrella ·  
R. W. Noda · M. L. F. Simeone · R. E. Schaffert ·  
J. V. de Magalhães · C. M. B. Damasceno (✉)  
Embrapa Maize and Sorghum, Sete Lagoas, Minas Gerais  
35701-970, Brazil  
e-mail: cynthia.damasceno@embrapa.br

**Keywords** Bioenergy · Predictive models · Accuracy ·  
Genotyping by sequencing · Functional enrichment

## Introduction

Increasing concerns about environmental issues have aroused global interest in the use of alternative sources for energy production. The use of plant biomass emerges as a viable alternative for the generation of biofuels (Rooney et al. 2007; Morris et al. 2013). Different organic materials have been tested, including high-biomass

sorghum (*Sorghum bicolor* L. Moench), which has potential to become an important crop for bioenergy production. This is due to its high biomass content, low water and fertilizer requirements, well-established production systems, tolerance to drought and heat, and high genetic diversity (Murray et al. 2008; Calviño and Messing 2012; Cotton et al. 2013; Brenton et al. 2016). High-biomass sorghum can be used for cellulosic ethanol or bioelectricity production. Cellulosic ethanol, or second generation ethanol, is produced from the hydrolysis of plant biomass into simple sugars, which in turn can be used for ethanol production by fermentation (Sticklen 2008; Zheng et al. 2009; Mandegari et al. 2017). The main focus of bioenergy-targeted breeding programs is to maximize biomass production per land unit, without increasing the energy input, in order to minimize the use of cultivable area. Breeders also aim at modifying cell wall polymers in order to facilitate the subsequent industrial processes. Modifying cell wall composition by, for example, reducing lignin content, may increase the efficiency at which biomass is converted into ethanol (Vermerris et al. 2007; Edmé et al. 2017).

Sorghum is a diploid species ( $2n=20$ ) with a relatively small genome, around 700 Mbp (Paterson et al. 2009), for which a reference sequence is already available. This resource facilitates breeding efforts involving molecular markers, because it is possible to confirm previous results and use genome annotations to make inferences concerning potential candidate genes, for example. Sorghum biomass breeding programs can take advantage of methods that reduce the time required to complete a breeding cycle and enable early and efficient selection of superior genotypes. Genomic selection has great potential to attain these goals. This strategy was first proposed by Meuwissen et al. (2001) to increase the efficiency of marker-assisted selection and accelerate the breeding cycle. The method uses markers distributed across the genome to predict the breeding value of individuals. Genomic selection was first applied in an animal breeding context, due to high cost of phenotyping and the impossibility of using replicates (Piyasatian et al. 2006; Solberg et al. 2006; Schaeffer 2006; Dekkers 2007; Long et al. 2007; Lee et al. 2008; Legarra et al. 2008; VanRaden et al. 2009). Later, it also attracted the attention of plant breeders (Bernardo and Yu 2007; Heffner et al. 2009; Jannink et al. 2010). Simulation and empirical studies performed

in various crops showed the superiority in terms of accuracy of genomic selection over traditional marker-assisted selection and selection based only on pedigree information (Bernardo and Yu 2007; Bernardo 2009, 2014a; Zhong et al. 2009; Lorenzana and Bernardo 2009; Mayor and Bernardo 2009; Crossa et al. 2010b, 2013, 2014; Grattapaglia and Resende 2010; Zhao et al. 2013).

In recent years, the development of next generation sequencing allowed genome-wide genotyping at lower costs. The genotyping-by-sequencing (GBS) technique is a multiplex system that allows the simultaneous identification of thousands of single-nucleotide polymorphisms (SNPs) and genotyping of the entire population of interest. This strategy has been used for a variety of species, such as barley, wheat, maize, rice, grapes, cocoa, sorghum, bean, soybean, cassava, cranberry, grass, sunflower, and oil palm (Elshire et al. 2011; Hansey et al. 2012; Poland et al. 2012; Sonah et al. 2013; Lu et al. 2013; Morris et al. 2013; Pootakham et al. 2015; Bredeson et al. 2016; Celik et al. 2016; Covarrubias-Pazaran et al. 2016; McAllister and Miller 2016). Due to the large number of markers, this technique is suitable for genomic selection (Poland and Rife 2012).

Various statistical models have been proposed for use in genomic selection (Meuwissen et al. 2001; Gianola et al. 2003; Park and Casella 2008; Habier et al. 2011). The main distinction between these models is the assumption about the underlying distribution of quantitative trait loci (QTL) effects. Due to particularities of the genetic architecture of different quantitative traits (Jiang and Zeng 1995; Zeng et al. 1999; Zeng 2001), distinct genomic selection models may be suitable for different phenotypic traits. Genomic selection models estimate the effects of individual markers and use information simultaneously from all markers available to estimate breeding values, without testing for individual marker effects; the aim is the selection directly applied to plant breeding (Bernardo and Yu 2007; Dekkers 2007; Goddard and Hayes 2007; Habier et al. 2007). Nevertheless, it is expected that markers within genes and/or with particular functional roles present effects of higher magnitude. Recently, novel models that exploit prior biological information in the analysis were proposed (Edwards et al. 2016; MacLeod et al. 2016). The predicted effects of markers in genomic selection studies can thus be used for functional enrichment analysis.

This work aimed to predict genomic breeding values of a high-biomass sorghum panel for bioenergy production. In addition, we investigated the potential use of functional enrichment analysis on marker-predicted effects for revealing important biological pathways involved in controlling quantitative traits related to biomass production and quality.

## Materials and methods

### Plant material

In this study, we analyzed a panel of 200 sorghum genotypes from Embrapa Maize and Sorghum. The panel is divided into two equally sized sub-panels. The 100 lines of sub-panel I are composed of materials from Embrapa germplasm bank and core collections from CIRAD and ICRISAT, consisting of 50 lines of high-biomass sorghum and 50 lines of saccharine sorghum. The remaining 100 lines of sub-panel II consist of high-biomass sorghum derived from Embrapa germplasm bank, originated mostly from accessions from the USA and some from Africa. These were added to Embrapa's breeding program and later to the panel and were characterized by lower genetic variability, higher sensitivity to photoperiod, and high levels of cellulose when compared to genotypes of sub-panel I. Each sorghum line in the panel is identified in Supplementary Table 1.

### Molecular data

The 200 lines were genotyped using the GBS technology. We applied the standard GBS protocol (Elshire et al. 2011) with restriction enzyme ApeKI and 96-plex approach. We then used the BWA (Li and Durbin 2009) to align tags to the reference genome of *Sorghum bicolor* (v2.1) and TASSEL-GBS (Glaubitz et al. 2014) pipeline for SNP calling.

We initially assessed the quality of these genotypic data with the following statistics: frequency of heterozygous genotypes for each SNP, minor allele frequency (MAF), and frequency of missing data for each genotype and each SNP. Missing marker values were then imputed using the NPUTE software (Roberts et al. 2007). After the imputation, we discarded markers with MAF lower than 5% and recoded the genotypes, assigning a value of -1 or 1 to the two possible homozygote genotypes.

### Population structure

We performed principal component analysis (PCA) to evaluate population structuring in the panel, using the R package *pcaMethods* (Stacklies et al. 2007). We also inferred population structure with STRUCTURE 2.3.4 (Pritchard et al. 2000). To warrant the use of unlinked markers, the genetic data were pruned to remove SNPs with an  $r^2$  value higher than 0.20, in sliding windows of 2 Mbp, using the SPNRelate R package (Zheng et al. 2012). Using a threshold value of 0.2 eliminates a large degree of redundancy in the data and reduces the influence of chromosomal artifacts (Laurie et al. 2010). The linkage disequilibrium pruning step left 23,117 SNPs, which were used to perform the structure analysis. We tested number of populations ( $K$ ) varying from one to ten, running a series of ten independent runs for each value of  $K$ . Each run consisted of a burn-in period of 100,000 and 200,000 MCMC iterations. For the choice of the most likely value of  $K$ , we used the  $\Delta K$  method, described by Evanno et al. (2005) and implemented in the Structure harvester software (Dent and Bridgett 2012). The most likely value of  $K$  was used to obtain conditional population membership coefficients of each individual.

### Phenotypic data

The 100 lines of sub-panel I were evaluated phenotypically for two years (2011 and 2012) and the 100 lines of sub-panel II for one year (2011), in Sete Lagoas, Minas Gerais State, Brazil. Summary statistics of these data are presented in supplementary material (Supplementary Tables 2 and 3; Supplementary Figs. 1 and 2). The experimental design consisted of a lattice (10 × 10) with three replicates. Plots were composed by 5 m lines spaced at 0.7 m, and showing nine plants per linear meter density. We evaluated the following phenotypic traits: days to flowering, number of days from seeding to the beginning of the pollen release in 50% of the plants in each plot; plant height, average height in meters of all plants in each plot, measured from the soil surface to the apex of the panicle; fresh matter yield (FMY) and dry matter yield (DMY), given in kg/plot by weighing all plants (whole) of each plot, harvested at grain physiological maturity, and then converted to t ha<sup>-1</sup>. To determine the dry matter, we took samples of fresh matter from the plot, which were incubated in an oven at 65 °C for 72 h or until sample weight was constant.

Subsequently, by the difference between the dry and fresh weights, we obtained the percentage of dry matter for each plot. Additionally, biomass samples were characterized by determining acid detergent fiber (ADF), neutral detergent fiber (NDF), cellulose, hemicellulose and lignin (acid detergent lignin), according to Van Soest et al. (1991), and their values are presented as a percentage of dry matter weight.

### Phenotypic analyses

We initially fitted appropriate mixed models for the phenotypic data of sub-panel I and sub-panel II. The statistical model for each trait of sub-panel I was as follows:

$$y_{ijkm} = \mu + \beta s_{ikm} + \gamma d_{ikm} + a_m + r_{km} + b_{jkm} + g_{im} + \varepsilon_{ijkm} \quad (1)$$

where  $y_{ijkm}$  is the phenotype of the  $i$ th genotype of block  $j$ , in replicate  $k$  and year  $m$ ;  $\mu$  is the intercept;  $s_{ikm}$  is the crop stand of the  $i$ th genotype, in replicate  $k$ , year  $m$ , corresponding to the count of plants that were effectively evaluated during the harvest period, and  $\beta$  is the corresponding fixed effect;  $d_{ikm}$  is a covariate representing the number of days to flowering for the  $i$ th genotype, in replicate  $k$ , year  $m$ , and  $\gamma$  is the corresponding fixed effect;  $a_m$  is the fixed effect of year  $m$ ;  $r_{km}$  is the random effect of replicate  $k$  in year  $m$ ;  $b_{jkm}$  is the random effect of block  $j$ , in replicate  $k$ , in year  $m$ ;  $g_{im}$  is the random effect of the  $i$ th genotype in year  $m$ ; and  $\varepsilon_{ijkm}$  is a random non-genetic effect. The correction for the effect of days to flowering aimed to eliminate the influence of early or late flowering on other traits. Particularly for biomass production, this allows selecting genotypes that contribute with high biomass alleles independently of their flowering behavior. This correction was included when fitting the model for all traits except days to flowering.

The statistical model for each trait of sub-panel II can be indicated by the following:

$$y_{ijk} = \mu + \beta s_{ik} + \gamma d_{ik} + r_k + b_{jk} + g_i + \varepsilon_{ijk} \quad (2)$$

where  $y_{ijk}$  is the random phenotypic effect of the  $i$ th genotype of block  $j$ , in replicate  $k$ ;  $s_{ik}$  is the crop stand of the  $i$ th genotype in replicate  $k$ , and  $\beta$  is the corresponding fixed effect  $d_{ik}$  is a covariate representing the

number of days to flowering for the  $i$ th genotype, in replicate  $k$ , and  $\gamma$  is the corresponding fixed effect;  $r_k$  is the random effect of replicate  $k$ ;  $b_{jk}$  is the random effect of block  $j$ , in replicate  $k$ ;  $g_i$  is the random effect of the  $i$ th genotype and  $\varepsilon_{ijk}$  is a random non-genetic effect.

The model assumes that the random effect of genotype  $g_i$  follows a normal distribution with zero mean and variance  $\sigma_g^2$ . For the effects of replicate  $r_{km}$  or  $r_k$ , block  $b_{jkm}$  or  $b_{jk}$  and non-genetic effects  $\varepsilon_{ijkm}$  or  $\varepsilon_{ijk}$ , we fit different (co)variance structures, including the identity, diagonal, compound symmetry and unstructured models (Smith et al. 2001). The variance-covariance (VCOV) matrices for these purposes have been structured for convenient grouping factors: replicates were used as a grouping factor for the block effect, and blocks within replicates were used as a grouping factor for the residual effects. Year was also used as a grouping factor for all these effects in sub-panel I, including genotype, which allows us to model both the main genotype effect and the genotype by year interaction. We initially compared different models for the VCOV structure of replicate, block and genetic effects, using the Bayesian Information Criterion (BIC; Schwarz 1978). We then evaluated similar structures for the non-genetic effects. Finally, we used the best fitting VCOV structure model to obtain the best linear unbiased predictors (BLUPs) of the genotypes for each analyzed trait (Supplementary Tables 4 and 5).

Fitting model (1) with a compound symmetry structure for the genotype by year interaction allowed the trait heritability to be estimated for sub-panel I, with the following equation:

$$h^2 = \frac{\sigma_g^2}{\left(\sigma_g^2 + \frac{\sigma_{ga}^2}{m} + \frac{\sigma_e^2}{n_r m}\right)}$$

where  $n_r$  is the number of replicates,  $m$  is the number of years,  $\sigma_g^2$  is the genetic variance component,  $\sigma_{ga}^2$  is the genotype by year interaction variance component, and  $\sigma_e^2$  is the residual variance component. Similarly, for sub-panel II, we estimated heritability based on model (2) with the following equation:

$$h^2 = \frac{\sigma_g^2}{\left(\sigma_g^2 + \frac{\sigma_e^2}{n_r}\right)}$$

where terms are as previously defined. All the analyses were performed using the software GenStat, version 16 (Payne et al. 2013).

#### Fitting of genomic selection models

We used the R packages BGLR (Pérez and de Los Campos 2014) to fit the Bayesian models Bayes A, Bayes B, Bayes C $\pi$ , Bayes Lasso and Bayes RR, and rrBLUP (Endelman 2011) to fit the random regression best linear unbiased predictor (RRBLUP) model. We used 40,000 iterations for Bayesian models, with 20,000 discarded as burn-in, and assumed default parameters for prior models. Our analyses used a cross-validation procedure to evaluate the ability of a model to predict breeding values. To this end, the 200 genotypes were divided into ten mutually exclusive groups, each containing 20 genotypes. For each cross-validation set, we began by fitting the genomic selection models on a training set of 180 genotypes, to estimate marker effects based on genotypic and phenotypic information. These marker effects then provided estimates of the breeding values of the remaining 20 individuals, based only on genotypic information—genomic estimated breeding value (GEBV). Finally, the correlation between the GEBVs and the estimated breeding values, i.e., the BLUPs obtained in the phenotypic analyses, provided estimates of the predictive abilities of the genomic selection models.

We also wanted to investigate the application of genomic selection models, trained on sets of selected individuals and years, to predict the behavior of genotypes across populations and/or years. The models were thus used for prediction across sub-panels, that is, the 100 genotypes of sub-panel I were used as training set and the 100 genotypes of sub-panel II as the test set. First, we used estimated breeding values for sub-panel I based on the combined data from two years. We also investigated the use of data for each year, separately, to assess the prediction across sub-panels and across sub-panels and years, respectively. Finally, we investigated the performance of genomic selection across years, training the models for sub-panel I in the first year to predict the behavior of the same genotypes in the second year.

#### Effect of marker density

We evaluated the effect of marker density on the efficiency of genomic selection. Multiple scenarios with

reduced numbers of markers were delineated and the RRBLUP model was fitted to each of them, providing estimates of predictive abilities. Starting from the complete set of 258,220 markers, we randomly removed half of the markers at each step, down to a minimum of 16. From the complete set of markers, ten random subsets were obtained for each tested marker density. We only used the RRBLUP model due to its lower computational requirements compared to the other models.

We also evaluated the consequences of removing markers of small effects. For this purpose, we initially fitted the RRBLUP model to all markers. Then, half of the markers with effects of greater magnitude was selected and used again to fit the RRBLUP model. We did this successively until a minimum of 16 markers, always choosing the markers of higher (absolute) effects (Supplementary Fig. 3). We did this selection of markers of larger effects using a ten-fold cross validation strategy as previously described, to avoid bias in the selection of markers.

#### Functional enrichment

Given a set of predicted marker effects, we wanted to test the hypothesis that some SNPs with particular biological functions had higher effects on the estimation of breeding values for each trait. To that end, we performed functional enrichment analysis of these marker effects. The sorghum genome obtained from the Phytozome platform (Paterson et al. 2009; Goodstein et al. 2012) includes predicted gene models and annotation of gene ontology (GO) terms for each predicted gene product (Paterson et al. 2009). This functional classification provides an ontology of terms representing the biological properties of a gene product (Ashburner et al. 2000). We assigned each SNP located inside a gene product all GO terms associated with that gene. All markers matching a given GO term found in this data set formed a distinct GO cluster. We then applied the Kolmogorov–Smirnov test (Frank and Massey 1951) to compare the distribution of the effects of all markers in a GO cluster with the distribution of the effects of all the remaining markers, individually for each trait. Our aim was to detect GO terms in which the marker effects were on average higher than the effects of the set of all markers. For this reason we chose a unilateral significance test to compare the absolute values of marker effects. Type I error control for multiple tests was done with the false discovery rate (FDR) correction (Benjamini and

Hochberg 1995). GO terms with adjusted  $p$  value  $< 0.01$  were deemed significant. We conducted the functional enrichment analysis separately for markers effects predicted with all genomic selection models.

## Results

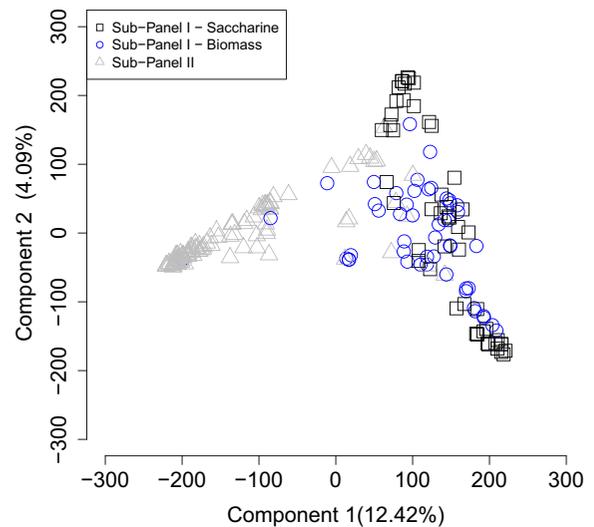
### Genotypic data

Genotyping of the 200 high-biomass sorghum genotypes generated a total of 1,024,892 SNPs, distributed along the ten chromosomes (Supplementary Fig. 4). The proportion of heterozygous genotypes per SNP ranged from 0 to 95%, with mean of 4.2% and median of 2.2% (Supplementary Fig. 5). These loci with high heterozygosity likely corresponded to spurious polymorphisms from duplicated genomic regions (Glaubitz et al. 2014), artifacts which were removed during marker imputation.

An evaluation of the proportion of missing data showed that SNPs had between 0 and 99.5% of missing genotype calls, with mean of 34.6% and median of 28.5% (Supplementary Fig. 6). Because sorghum is a diploid, self-pollinating species with an available reference genome, imputation of missing data is greatly facilitated. After imputation was performed, the distribution of minor allele frequencies showed that most SNPs had MAF lower than 5% (Supplementary Fig. 7). In general, low MAF rates may represent sequencing errors, rare alleles, and low coverage. For this reason, we chose to filter SNPs with a minimum MAF value of 5%, which retained 258,220 SNPs, which were used for the fitting of genomic selection models.

### Population structure

Principal component analysis revealed structuring of genotypes between the two sub-panels. Differences in the first principal component reflected the separation of individuals from sub-panels I and II (Fig. 1). Interestingly, sub-panel I did not show distinctive structuring between saccharine and biomass genotypes, with the former being separated into two main groups, but with substantial overlapping of genotypes. These results are in agreement with those obtained by STRUCTURE, which indicated the best fit was for a value of  $K$  of two (Supplementary Fig. 8), with membership coefficients of individuals coinciding with the first PCA component (Supplementary Fig. 9).



**Fig. 1** Scatter plot of the two first principal component scores of 200 high-biomass sorghum genotypes. Sorghum genotypes belong to the Embrapa Maize and Sorghum germplasm bank and breeding program. Component scores were obtained from a Principal Component Analysis based on 258,220 SNP markers. Each solid circle represents a genotype, and the colors indicate the sub-panel it belongs

### Predictive ability

We observed through the estimated correlation between the breeding values predicted by genomic selection and the estimated breeding values that the highest predictive abilities were achieved for traits ADF, NDF, cellulose, and lignin, while traits hemicellulose and days to flowering had the lowest values (Table 1). In general, all predictive abilities obtained were high, ranging from 0.85 for NDF to 0.66 for days to flowering.

All tested genomic selection models yielded similar predictive abilities for each of the nine traits. Even though differences between models were modest, the RRBLUP model showed the best predictions overall, while the Bayes Lasso model showed the lowest predictive abilities. For example, for the trait plant height, the best and worst models provided values of 0.78 and 0.76, respectively.

We note that the trait days to flowering yielded the lowest predictive abilities, despite having heritabilities of 0.81 and 0.87, for sub-panel I and II, respectively. In contrast, hemicellulose showed lower heritabilities, 0.39 (sub-panel I) and 0.51 (sub-panel II), but yielded a predictive ability of 0.68.

To assess whether these accuracies were influenced by the population structure present between sub-panels,

**Table 1** Predictive abilities obtained from six genomic selection models applied to nine traits of the high-biomass sorghum panel of Embrapa Maize and Sorghum in the joint analysis. Values indicate

the correlation coefficient between the breeding values predicted by genomic selection models and the phenotypic breeding values

| Trait             | Heritability |              | Genomic selection model |        |         |        |            |        |
|-------------------|--------------|--------------|-------------------------|--------|---------|--------|------------|--------|
|                   | Sub-panel I  | Sub-panel II | BayesB                  | BayesA | BayesRR | BayesC | BayesLasso | RRBLUP |
| Plant height      | 0.96         | 0.83         | 0.77                    | 0.77   | 0.77    | 0.77   | 0.76       | 0.78   |
| Cellulose         | 0.78         | 0.89         | 0.83                    | 0.83   | 0.83    | 0.83   | 0.82       | 0.83   |
| ADF               | 0.83         | 0.86         | 0.83                    | 0.83   | 0.83    | 0.83   | 0.82       | 0.84   |
| NDF               | 0.76         | 0.88         | 0.84                    | 0.85   | 0.84    | 0.84   | 0.84       | 0.85   |
| Days to flowering | 0.81         | 0.87         | 0.64                    | 0.64   | 0.64    | 0.63   | 0.61       | 0.66   |
| Hemicellulose     | 0.39         | 0.51         | 0.68                    | 0.68   | 0.68    | 0.68   | 0.67       | 0.68   |
| Lignin            | 0.82         | 0.61         | 0.82                    | 0.82   | 0.82    | 0.82   | 0.82       | 0.82   |
| DMY               | 0.70         | 0.67         | 0.73                    | 0.73   | 0.73    | 0.73   | 0.72       | 0.74   |
| FMY               | 0.80         | 0.85         | 0.77                    | 0.77   | 0.77    | 0.77   | 0.76       | 0.77   |

ADF, fiber proportions in acid detergent; NDF, fiber proportions in neutral detergent; DMY, dry matter yield; FMY, fresh matter yield

we also performed a validation of predictive accuracy within each sub-panel. The predictive accuracy was high for all traits in both sub-panels, except for the trait hemicellulose in sub-panel II (Supplementary Table 6).

When we performed the prediction across sub-panels, using data from both years for sub-panel I, the predictive abilities achieved were considerably lower compared to those obtained in the analysis of the complete panel (Supplementary Table 7). For example, for the trait hemicellulose the predictive ability ranged from 0.68 in the analysis of the complete panel to  $-0.04$  in the analysis across sub-panels. In this situation we observed a stronger correlation between heritability and predictive abilities (Fig. 2). Interestingly, when we evaluated prediction across sub-panels (Supplementary Table 8), training the models only with data from the first year, we obtained predictive abilities higher in most traits, compared to those obtained in the analysis with both years. Indeed, the use of genomic selection models across sub-panels and years (Supplementary Table 9) resulted in the lowest predictive abilities. Finally, when analyzing the use of models across years for sub-panel I (Supplementary Table 10), observed predictive abilities were high indicating a better performance of genomic selection when the training and test populations were more closely related.

### Marker density

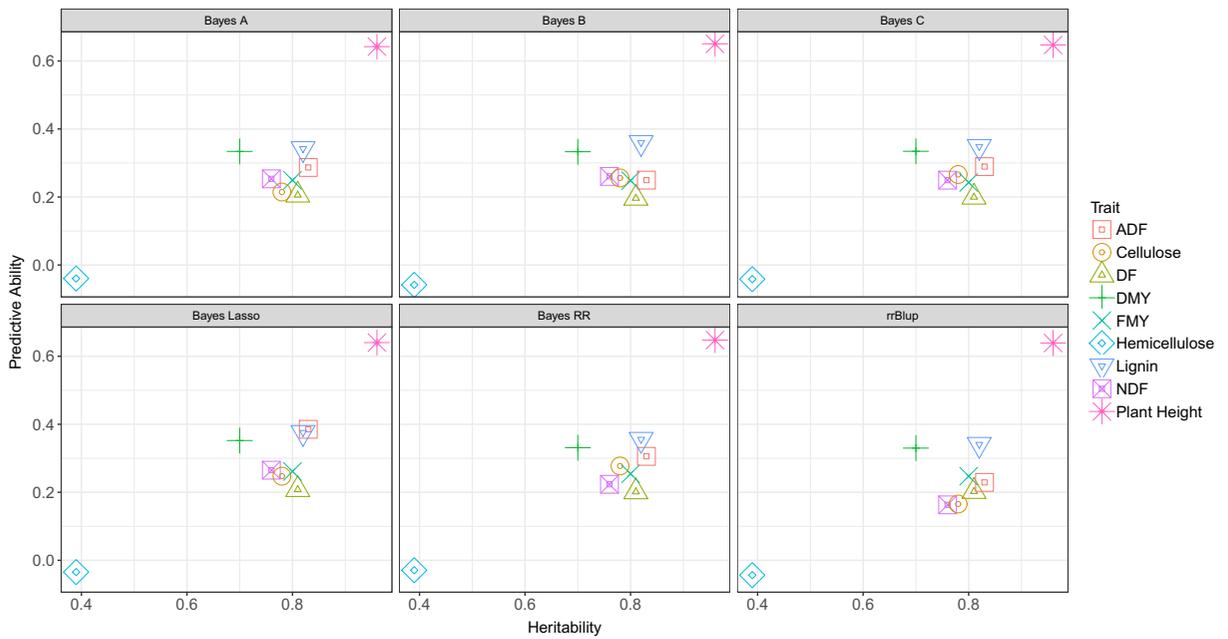
Effect of marker density on the predictive abilities of the model RRBLUP for each of the nine traits is shown in

Fig. 3. As the marker density increased, the predictive abilities also increased until reaching a plateau starting with roughly 2018 markers. The predictive abilities were maximum and the variance minimum when using the complete set of markers, for all traits. Conversely, the lowest marker density resulted in the minimum predictive abilities and maximum variance. For example, in the trait plant height, the predictive ability varied from 0.77 with 258,220 markers to 0.25 with 16 markers.

When we fitted the model using the markers with larger effects, we observed little influence on the predictive abilities compared with those obtained with the complete set of markers (Supplementary Fig. 3). Similarly to our observation based on the random selection of markers, predictive accuracies remained high even using one to eight thousand markers. For example, for the trait plant height the correlation varied from 0.80 to 0.78, with 258,220 and 505 markers, respectively. However, for reduced numbers of markers (i.e., less than 1009), the predictive abilities achieved were considerably lower. With the minimum of 16 markers the predictive abilities ranged from 0.53 for the trait hemicellulose to 0.76 for the trait NDF.

### Functional enrichment

Analysis of enrichment of marker effects predicted with different genomic selection models yielded different numbers of enriched GO terms (results not shown). Bayesian models resulted in small numbers of terms,

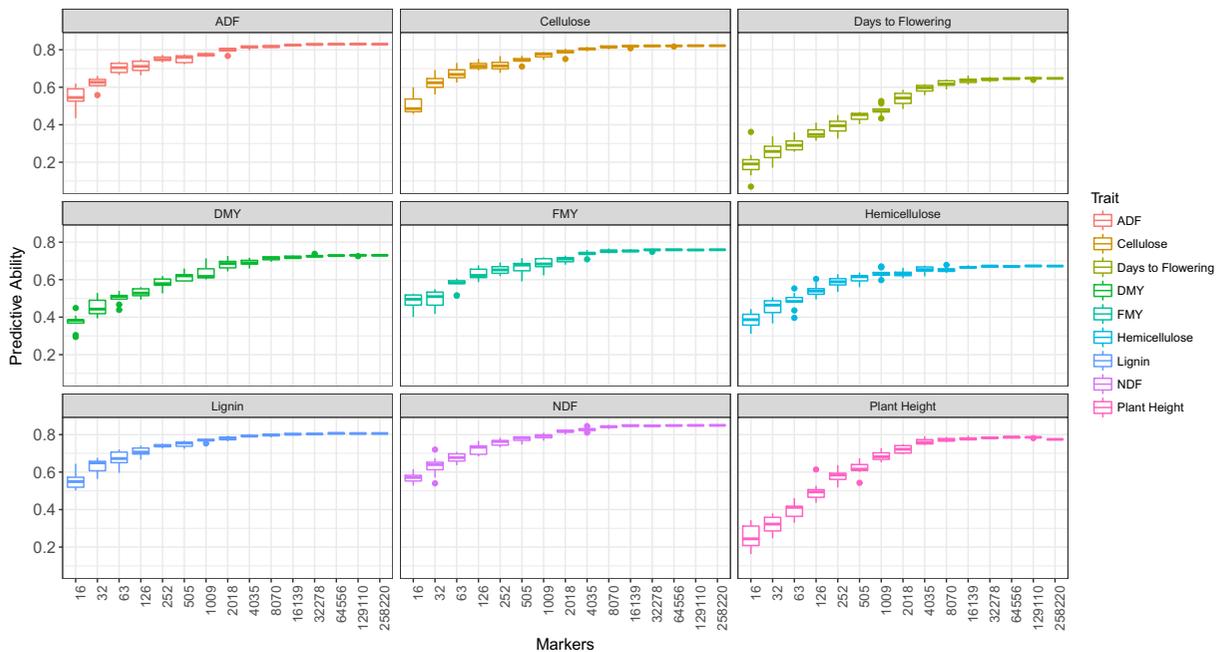


**Fig. 2** Relationship between trait heritability and predictive ability for different genomic selection models. Models Bayes A, Bayes B, Bayes C $\pi$ , Bayes Lasso, Bayes RR, and RRBLUP were

applied to nine traits of the high biomass sorghum panel of Embrapa Maize and Sorghum, for the prediction across sub-panels

which were also largely identified with the RRBLUP model. For this reason, we choose to only present the results from RRBLUP. We found a total of 1119 GO terms related to the SNPs present inside the predicted gene models. The number of SNPs associated with a

single GO term ranged from 1 to 14,625, with mean 214.9 and median 26. The number of GO terms associated with a single SNP ranged from 1 to 40, with mean 3.27 and median 3. The results of functional enrichment analyses based on the Kolmogorov-Smirnov test



**Fig. 3** Predictive abilities of the model RRBLUP as a function of marker density for nine traits of the high-biomass sorghum panel



Frisch 2014). However, it can still be relevant to compare different models, because successive application of genomic selection through breeding cycles may affect the outcome of models with different assumptions (Habier et al. 2007).

Phenotyping is very important in genomic selection studies, since it impacts all steps of the process, from the prediction of markers effects to the selection of superior genotypes. Thus, the precision with which phenotypic measurements are taken for each trait influences the predictive abilities of the models (Heslot et al. 2015). Days to flowering is an extremely important trait for breeding programs of high-biomass sorghum. Indeed, the genetic characterization of a regulatory system responsible for controlling the photoperiod and flowering time in sorghum enabled the development of high-biomass hybrids. These hybrids are highly sensitive to photoperiod and with late flowering, which exhibit prolonged vegetative growth and high biomass accumulation (Rooney and Aydin 1999; Rooney et al. 2007; Murphy et al. 2011). For this reason, we included a covariate to adjust for differences in days to flowering between genotypes.

The size and composition of the training population are important factors that can be manipulated by breeders (Habier et al. 2010; Rincent et al. 2012). The combination of training datasets maximizes the use of phenotypic and genotypic information available, generating large training populations and increasing the predictive ability (De Roos et al. 2009; Hayes et al. 2009; Asoro et al. 2011; Lorenz et al. 2012; Technow et al. 2013). In our study, we conducted joint analysis of the two sub-panels to ensure that a larger training population was used. Hence, the fact that the predictive abilities found (Table 1) were higher than those obtained in the cross analysis between sub-panels (Supplementary Table 6) can be partly explained by the size of the training population in each scenario. It is also expected that genetic predictions are more accurate for traits with high heritability (Combs and Bernardo 2013; Lorenz 2013). We observed such a trend in the analysis between sub-panels (Fig. 2). However, in the joint analysis we did not observe a strong trend between heritability and prediction accuracy (Table 1). Several authors showed that predictive abilities are also affected by effective population size, training population size, linkage disequilibrium, trait architecture, marker density, choice of predictive model and the relationship between training and breeding population (Daetwyler et al. 2008;

Grattapaglia and Resende 2010; Asoro et al. 2011; Nakaya and Isobe 2012; de los Campos et al. 2013). In this way, the fact that the prediction ability for plant height was not as high as its heritability could suggest that the predictive ability of this trait was affected by other factors. Similarly, the days to flowering trait had one of the lowest predictive abilities, despite showing high heritability. This lack of correlation between predictive ability and heritability is in agreement with other results in the literature (Grattapaglia and Resende 2010; Heffner et al. 2011).

One of the questions that arise in genomic selection studies is that the training and testing populations must be highly related to ensure an effective selection. The interaction of trait architecture and population structure plays an important role in creating a training population. Our principal component analysis (Fig. 1) showed that the structure of the 200 genotypes of the panel reflect their subdivision in two sub-panels. Sub-panel I has mostly saccharine and biomass genotypes from CIRAD and ICRISAT, while sub-panel II is mainly composed of genotypes from the Embrapa sorghum germplasm bank and breeding program. In the joint analysis, the similarity between genotypes in training and test populations was large. This can partially explain the fact that the predictive abilities found in the joint analysis (Table 1) were higher than those obtained in the cross analysis between sub-panels (Supplementary Table 7). Nonetheless, it is interesting to note that predictive abilities were still moderately high for plant height. In addition to the influence of population structure on accuracy, we observed a negative impact of genotype by year interaction on the use of genomic selection between sub-panels. Indeed, the predictive ability between sub-panels was higher when we only used data from the first year for sub-panel I, and inclusion of data from the second year reduced accuracy (Supplementary Tables 7 and 8).

The effects of marker density on the predictive abilities showed that the best predictions were obtained using the maximum marker density (Fig. 3). The median predictive abilities remained relatively constant with reduced numbers of markers, down to roughly four thousand markers. From a practical standpoint, this indicates that a reduced number of markers explained most of the genetic variation, opening new perspectives for the use of a relatively small subset of SNPs in sorghum, as is done for other plant and animal species, by constructing genotyping chips (Matukumalli et al. 2009; Yu et al. 2013; Wang et al. 2014). In any case, this

may not be true for other breeding scenarios, and using the complete set of markers resulted in the highest accuracies (Meuwissen and Goddard 2010).

Genomic selection has revolutionized the use of marker-assisted selection in plant breeding, mainly due to its distinct approach when compared to QTL mapping and genome-wide association studies (GWAS). Standard QTL mapping aims to map chromosomal regions affecting phenotypic traits of interest, thus enabling the use of markers linked to these regions (Bernardo 2008; Lorenz et al. 2011). However, the use of QTL mapping in breeding programs is limited by the fact that the commonly used bi-parental populations have applications that may be conditioned to the specific population under study. Besides that, the statistical models used are unsuitable for breeding of polygenic traits, which are controlled by many loci of small effect (Meuwissen et al. 2001; Goddard and Hayes 2007; Heffner et al. 2009). The main objective of GWAS is also the identification of chromosomal regions associated with a particular trait, using a diversity panel instead of a breeding population (Ingvarsson and Street 2011; Huang and Han 2014). In contrast, genomic selection does not intend to test for the significance of genes and/or individual markers. It instead leverages information from all available genome-wide markers. This makes this methodology directly applicable to plant breeding (Bernardo and Yu 2007; Dekkers 2007; Goddard and Hayes 2007; Habier et al. 2007).

Nevertheless, we expect that markers located near genes responsible for certain biological functions present effects of greater magnitude. Thus, the predicted effects of markers in genomic selection studies can be used for functional enrichment analysis to identify particularly important functional groups. We stress that this strategy only considered markers located within genes, but our results indicate that 28.4% of the SNPs were located in functionally annotated genes. Our data revealed several interesting associations for the several traits evaluated, which should be further investigated. As an example, the detection of GO terms related to microtubule-based movement for the trait cellulose (Fig. 4) is possibly connected with the fact that the deposition of cellulose is guided by microtubules that are adjacent or directly connected to a synthesis complex (Delmer and Amor 1995). Besides, in agreement with published studies (Gillmor et al. 2005; Ben-Tov et al. 2015), GO terms related to GPI anchor metabolic process are expected to be associated with cellulose deposition. We also

observed several carbohydrate-related GO terms for hemicellulose, such as carbohydrate transport, Golgi membrane, GPI-anchor metabolism, and cell wall modification. After hemicellulose is synthesized in the Golgi complex, it is transported to the plasma membrane, so the identification of these GO terms is expected (Pauly et al. 2013). Lignin biosynthesis is initiated in the cytosol with the synthesis of glycosylated monolignols from phenylalanine, an aromatic amino acid. Interestingly, the GO term related to the aromatic amino acid family biosynthetic process is among those enriched for lignin (Boerjan et al. 2003). For the traits days to flowering and plant height, we observed the enriched GO term autophagy. In plant cells, autophagy plays roles in recycling of proteins and metabolites including lipids, and is involved in many physiological processes, such as abiotic and biotic stress response. In addition, autophagy has particular importance on male reproductive development during pollen maturation (Hanamata et al. 2014). Overall, multiple terms related to the biosynthesis of macromolecules were detected for several of the biomass related traits evaluated. This indicates that metabolic processes involving primary metabolites can be important for predicting breeding values. In that case, the selection of genotypes according to their GEBVs may exert stronger selective pressure on these SNPs, because their effects were of greater magnitude. By using this functional enrichment approach, we attempted to couple the application of genomic selection for breeding purposes with the association of particular functional classes of markers with the phenotypic traits. Brenton et al. (2016) performed genome-wide association analysis in a sorghum panel composed of sweet and biomass types and identified potential genes that could lead to bioenergy sorghum improvement. Interestingly, these authors identified a region on chromosome 6 associated with NDF, that had two genes coding for cellulase enzymes, *Sobic.006G122200* and *Sobic.006G122300*. These gene products are responsible for hydrolyzing glycosidic bonds in complex carbohydrates. We also found terms related to the hydrolysis of O-glycosyl compounds (GO:0004553) and carbohydrate metabolic process (GO:0005975) in the enriched gene ontology terms for the trait NDF.

One of the main advantages of genomic selection in breeding programs is the reduced time needed to develop new materials. Genomic selection can reduce the breeding cycle through early prediction of phenotypic performance of a set of genotypes for various traits of

interest (Meuwissen et al. 2001; Bernardo and Yu 2007; Bernardo 2008; Crossa et al. 2011). It can also be used to predict phenotypic performance of genotypes for traits of difficult evaluation, such as those related to biomass composition (ADF, NDF, and lignin), which require expensive and laborious phenotypic evaluations. Genomic selection studies in plants have been based on breeding populations, real or simulated (Bernardo and Yu 2007; Crossa et al. 2010a; Zhao et al. 2013). However, in this study we used collections of genotypes with large genetic variability, which constitute diversity panels. Using these panels, we aimed to reduce the time required to select genotypes of biomass sorghum in early stages of Embrapa breeding program, through prediction based on genotypes already selected in other breeding programs (Supplementary Table 6). We achieved predictive abilities in the order of 0.39, 0.38, 0.35, 0.28, 0.26, 0.26, and 0.21 for the traits ADF, lignin, DMY, cellulose, NDF, FMY, and days to flowering, respectively, while for plant height we observed a predictive ability of 0.65. However, for the trait hemicellulose, the predictive abilities found were low and negative. This might be due to the compositional analysis method used to measure hemicellulose content, which is faster and less costly than chromatography-based methods, but also less accurate.

Currently, several studies that apply genomic selection in plant species have been developed (Bernardo and Yu 2007; Bernardo 2009, 2014b; Heffner et al. 2009; Mayor and Bernardo 2009; Jannink et al. 2010; Grattapaglia and Resende 2010; Poland et al. 2012; Zhao et al. 2013; Crossa et al. 2013, 2014; Zhang et al. 2014). Using biomass sorghum, Yu et al. (2016), showed the potential use of genomic selection to improve the process of germplasm evaluation in global gene banks. This innovative way to apply this strategy could facilitate downstream breeding and genetic research. With a different approach, our work shows that genomic selection can be successfully applied directly in biomass sorghum breeding programs, which have the potential to help sorghum become an important bioenergy feedstock in Brazil. Yu et al. (2016) applied genomic selection to study eight biomass related traits, including dry biomass yield and plant height. Using a cross-validation scheme they found predictive abilities that ranged from 0.35 to 0.78. Similarly, in our work we found high predictive abilities for the several traits evaluated, contributing to an early and efficient selection of the best genotypes. The models of genomic selection

used herein yielded satisfactory results, which are directly applicable to breeding and potentially able to reduce the time required for the launching of new cultivars of biomass sorghum, increasing the potential for this important bioenergy crop. Finally, our functional enrichment analysis attempts to show that, although genomic selection is not primarily focused on identifying and testing markers associated with phenotypes, its results can help in understanding the biological processes involved in the expression of quantitative traits.

**Acknowledgments** AAO received a fellowship from grant #2013/25132-7, São Paulo Research Foundation (FAPESP), and VFS received a fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). This research was supported by grant #2015/22993-7, São Paulo Research Foundation (FAPESP) to GRAM; Fundação de Amparo à Pesquisa do Estado de Minas Gerais, FAPEMIG (grant number APQ-02237-13), and Embrapa (Brazilian Agricultural Research Corporation).

#### Compliance with ethical standards

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

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

#### References

- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. *Nat Genet* 25:25–29
- Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink JL (2011) Accuracy and training population design for genomic selection on quantitative traits in elite north American oats. *Plant Genome J* 4:132–144. <https://doi.org/10.3835/plantgenome2011.02.0007>
- Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, Vaillancourt B, Buell CR, Kaeppler SM, de Leon N (2013) Marker density and read depth for genotyping populations using genotyping-by-sequencing. *Genetics* 193: 1073–1081. <https://doi.org/10.1534/genetics.112.147710>

- Ben-Tov D, Abraham Y, Stav S, Thompson K, Loraine A, Elbaum R, Souza A, Paulv M, Kibier JJ, Harpaz-Saad S (2015) COBRA-LIKE 2, a member of the GPI-anchored COBRA-LIKE family, plays a role in cellulose deposition in Arabidopsis seed coat mucilage secretory cells. *Plant Physiol* 167. <https://doi.org/10.1104/pp.114.240671>
- Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 57:289–300
- Bernardo R (2009) Genomewide selection for rapid introgression of exotic germplasm in maize. *Crop Sci* 49:419. <https://doi.org/10.2135/cropsci2008.08.0452>
- Bernardo R (2014a) Genomewide selection when major genes are known. *Crop Sci* 54:68. <https://doi.org/10.2135/cropsci2013.05.0315>
- Bernardo R (2014b) Essentials of plant breeding. Stemma Press, Woodbury
- Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Sci* 48: 1649. <https://doi.org/10.2135/cropsci2008.03.0131>
- Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. *Crop Sci* 47:1082–1090. <https://doi.org/10.2135/cropsci2006.11.0690>
- Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. *Annu Rev Plant Biol* 54:519–546. <https://doi.org/10.1146/annurev.arplant.54.031902.134938>
- Bredeson JV, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, Grimwood J, Schmutz J, Rabbi IY, Egesi C, Nauluvula P, Lebot V, Ndunguru J, Mkamilo G, Bart RS, Setter TL, Gleadow RM, Kulakow P, Ferguson ME, Rounsley S, Rokhsar DS (2016) Sequencing wild and cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nat Biotechnol* 34: 562–570. <https://doi.org/10.1038/nbt.3535>
- Brenton ZW, Cooper EA, Myers MT, Boyles RE, Shakoore N, Zielinski KJ, Rauh BL, Bridges WC, Morris GP, Kresovich S (2016) A genomic resource for the development, improvement, and exploitation of sorghum for bioenergy. 204:21–33. <https://doi.org/10.1534/genetics.115.183947>
- Calviño M, Messing J (2012) Sweet sorghum as a model system for bioenergy crops. *Curr Opin Biotechnol* 23:323–329. <https://doi.org/10.1016/j.copbio.2011.12.002>
- Celik I, Bodur S, Fray A, Doganlar S (2016) Genome-wide SNP discovery and genetic linkage map construction in sunflower (*Helianthus annuus* L.) using a genotyping by sequencing (GBS) approach. *Mol Breed*. doi: <https://doi.org/10.1007/s11032-016-0558-8>
- Combs E, Bernardo R (2013) Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *Plant Genome* 6:1–7. <https://doi.org/10.3835/plantgenome2012.11.0030>
- Cotton J, Acosta-Martínez V, Moore-Kucera J, Burow G (2013) Early changes due to sorghum biofuel cropping systems in soil microbial communities and metabolic functioning. *Biol Fertil Soils* 49:403–413. <https://doi.org/10.1007/s00374-012-0732-z>
- Covarrubias-Pazarán G, Díaz-García L, Schlautman B, Deutsch J, Salazar W, Hernández-Ochoa M, Grygleski E, Steffan S, Iorizzo M, Polashock J, Vorsa N, Zalapa J (2016) Exploiting genotyping by sequencing to characterize the genomic structure of the American cranberry through high-density linkage mapping. *BMC Genomics* 17:451. <https://doi.org/10.1186/s12864-016-2802-3>
- Crossa J, de los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010a) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186:713–724. <https://doi.org/10.1534/genetics.110.118521>
- Crossa J, Pérez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokoshol C (2010b) Genomic prediction of quantitative traits in plant breeding using molecular markers and pedigree. In: *Cosmos*. pp 1–33
- Crossa J, Pérez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokoshol C (2011) Genomic selection and prediction in plant breeding. *J Crop Improv* 25:239–261. <https://doi.org/10.1080/15427528.2011.558767>
- Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E, Jannink JL, Lopez Cruz MA, Babu R (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3* 3:1903–26. <https://doi.org/10.1534/g3.113.008227>
- Crossa J, Pérez P, Hickey J, Burgueño J, Ormella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* (Edinb) 112:48–60. <https://doi.org/10.1038/hdy.2013.16>
- Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3:e3395. <https://doi.org/10.1371/journal.pone.0003395>
- Dassonneville R, Brøndum RF, Druet T, Fritz S, Guillaume F, Guldbandsen B, Lund MS, Ducrocq V, Su G (2011) Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. *J Dairy Sci* 94:3679–3686. <https://doi.org/10.3168/jds.2011-4299>
- Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* 12:499–510. <https://doi.org/10.1038/nrg3012>
- de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193: 327–345. <https://doi.org/10.1534/genetics.112.143313>
- De Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. *Genetics* 183:1545–1553. <https://doi.org/10.1534/genetics.109.104935>
- Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. *J Anim Breed Genet* 124:331–341
- Delmer DP, Amor Y (1995) Cellulose biosynthesis. *Plant Cell* 7: 987–1000. <https://doi.org/10.1146/annurev.pp.38.060187.001355>
- Dent AE, Bridgett M vH (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* 4:359–361. <https://doi.org/10.1007/s12686-011-9548-7>
- Edmé S, Mitchell R, Sarath G (2017) Genetic parameters and prediction of breeding values in switchgrass bred for

- bioenergy. *Crop Sci* 57:0. doi: <https://doi.org/10.2135/cropsci2016.09.0770>
- Edwards SM, Sørensen IF, Sarup P, Mackay TFC, Sørensen P (2016) Genomic prediction for quantitative traits is improved by mapping variants to gene ontology categories in *Drosophila melanogaster*. *Genetics* 203:1871–1883. <https://doi.org/10.1534/genetics.116.187161>
- 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* 6:1–10. <https://doi.org/10.1371/journal.pone.0019379>
- Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rBLUP. *Plant Genome J* 4: 250. <https://doi.org/10.3835/plantgenome2011.08.0024>
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14:2611–2620. <https://doi.org/10.1111/j.1365-294X.2005.02553.x>
- Frank J, Massey J (1951) Journal of the American statistical. *J Am Stat Assoc* 46:68–78
- Gianola D, Perez-Enciso M, Toro MA (2003) On marker-assisted prediction of genetic value: beyond the ridge. *Genetics* 163: 347–365
- Gillmor CS, Lukowitz W, Brininstool G, Sedbrook JC, Hamann T, Poindexter P, Somerville C (2005) Glycosylphosphatidylinositol-anchored proteins are required for cell wall synthesis and morphogenesis in *Arabidopsis*. *Plant Cell* 17:1128–1140. <https://doi.org/10.1105/tpc.105.031815.a>
- 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>
- Goddard ME, Hayes BJ (2007) Genomic selection. *J Anim Breed Genet* 124:323–330. <https://doi.org/10.1111/j.1439-0388.2007.00702.x>
- Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40:1178–1186. <https://doi.org/10.1093/nar/gkr944>
- Grattapaglia D, Resende MDV (2010) Genomic selection in forest tree breeding. *Tree Genet Genomes* 7:241–255. <https://doi.org/10.1007/s11295-010-0328-4>
- Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177:2389–2397. <https://doi.org/10.1534/genetics.107.081190>
- Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low-density marker panels. *Genetics* 182:343–353. <https://doi.org/10.1534/genetics.108.100289>
- Habier D, Tetens J, Seefried F-R, Lichtner P, Thaller G (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution* 42:5.
- Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186. <https://doi.org/10.1186/1471-2105-12-186>
- Hanamata S, Kurusu T, Kuchitsu K (2014) Roles of autophagy in male reproductive development in plants. *Front Plant Sci* 5: 457. <https://doi.org/10.3389/fpls.2014.00457>
- Hansey CN, Vaillancourt B, Sekhon RS, de Leon N, Kaeppler SM, Buell CR (2012) Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. *PLoS One* 7:1–10. <https://doi.org/10.1371/journal.pone.0033071>
- Hayes BJ, Bowman PJ, Chamberlain A, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 92:433–443. <https://doi.org/10.3168/jds.2008-1646>
- Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. *Crop Sci* 49(1). <https://doi.org/10.2135/cropsci2008.08.0512>
- Heffner EL, Jannink J-L, Iwata H et al (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci* 51:2597–2606. <https://doi.org/10.2135/cropsci2011.05.0253>
- Heslot N, Jannink J-L, Sorrells ME (2015) Perspectives for genomic selection applications and research in plants. *Crop Sci* 55. <https://doi.org/10.2135/cropsci2014.03.0249>
- Hofheinz N, Frisch M (2014) Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. *G3 (Bethesda)* 4:539–546. <https://doi.org/10.1534/g3.113.010025>
- Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. *Annu Rev Plant Biol* 65: 531–551. <https://doi.org/10.1146/annurev-arplant-050213-035715>
- Ingvarsson PK, Street NR (2011) Association genetics of complex traits in plants. *New Phytol* 189:909–922. <https://doi.org/10.1111/j.1469-8137.2010.03593.x>
- Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics* 9:166–177. <https://doi.org/10.1093/bfpg/elq001>
- Jiang C, Zeng ZB (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140:1111–1127. <https://doi.org/10.1111/j.1567-1364.2011.00777.x>
- Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. *NIH Public Access Genet Epidemiol* 34:591–602. <https://doi.org/10.1002/gepi.20516>
- Lee SH, Van Der Werf JHJ, Hayes BJ, Goddard ME, Visscher PM (2008) Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genet* 4:e1000231. <https://doi.org/10.1371/journal.pgen.1000231>
- Legarra A, Robert-Granié C, Manfredi E, Elsen JM (2008) Performance of genomic selection in mice. *Genetics* 180: 611–618. <https://doi.org/10.1534/genetics.108.088575>
- Li H, Durbin R, (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760
- Long N, Gianola D, Rosa GJM, Weigel KA, Avendaño S (2007) Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J Anim Breed Genet* 124:377–389. <https://doi.org/10.1111/j.1439-0388.2007.00694.x>

- Lorenz A J, Smith KP, Jannink JL (2012) Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. *Crop Sci* 52:1609–1621. <https://doi.org/10.2135/cropsci2011.09.0503>
- Lorenz AJ (2013) Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. *G3 (Bethesda)* 3:481–491. <https://doi.org/10.1534/g3.112.004911>
- Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink J-L (2011) Genomic selection in plant breeding: knowledge and prospects, 1st edn. Elsevier Inc.
- Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor Appl Genet* 120:151–161. <https://doi.org/10.1007/s00122-009-1166-3>
- 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* 9:1–14. <https://doi.org/10.1371/journal.pgen.1003215>
- MacLeod IM, Bowman PJ, Vander Jagt CJ, Haile-Mariam M, Kemper KE, Chamberlain AJ, Schrooten C, Hayes BJ, Goddard ME (2016) Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics* 17:144. <https://doi.org/10.1186/s12864-016-2443-6>
- Mandegari MA, Farzad S, Görgens JF (2017) Economic and environmental assessment of cellulosic ethanol production scenarios annexed to a typical sugar mill. *Bioresour Technol* 224:314–326
- Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TPL, Sonstegard TS, van Tassell CP (2009) Development and characterization of a high density SNP genotyping assay for cattle. *PLoS One* 4:e5350. <https://doi.org/10.1371/journal.pone.0005350>
- Mayor PJ, Bernardo R (2009) Genomewide selection and marker-assisted recurrent selection in doubled haploid versus F populations. *Crop Sci* 49:1719. <https://doi.org/10.2135/cropsci2008.10.0587>
- McAllister CA, Miller AJ (2016) Single nucleotide polymorphism discovery via genotyping by sequencing to assess population genetic structure and recurrent polyploidization in *Andropogon gerardii*. *Am J Bot* 103:1314–1325. <https://doi.org/10.3732/ajb.1600146>
- Meuwissen T, Goddard M (2010) Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics* 185:623–631. <https://doi.org/10.1534/genetics.110.116590>
- Meuwissen T, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, Riera-Lizarazu O, Brown PJ, Acharya CB, Mitchell SE, Harriman J, Glaubitz JC, Buckler ES, Kresovich S (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc Natl Acad Sci U S A* 110:453–458. <https://doi.org/10.1073/pnas.1215985110>
- Mulder H a, Calus MPL, Druet T, Schrooten C (2012) Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *J Dairy Sci* 95:876–889. doi: <https://doi.org/10.3168/jds.2011-4490>
- Murphy RL, Klein RR, Morishige DT, Brady JA, Rooney WL, Miller FR, Dugas DV, Klein PE, Mullet JE (2011) Coincident light and clock regulation of pseudoreponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. *Proc Natl Acad Sci U S A* 108:16469–16474. <https://doi.org/10.1073/pnas.1106212108>
- Murray SC, Sharma A, Rooney WL, Klein PE, Mullet JE, Mitchell SE, Kresovich S (2008) Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain nonstructural carbohydrates. *Crop Sci* 48:2165–2179. <https://doi.org/10.2135/cropsci2008.01.0016>
- Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? *Ann Bot* 110:1303–1316. <https://doi.org/10.1093/aob/mcs109>
- Park T, Casella G (2008) The Bayesian lasso. *J Am Stat Assoc* 103:681–686. <https://doi.org/10.1198/016214508000000337>
- Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberger G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otiillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature* 457:551–556. <https://doi.org/10.1038/nature07723>
- Pauly M, Gille S, Liu L, Mansoori N, de Souza A, Schultink A, Xiong G (2013) Hemicellulose biosynthesis. *Planta* 238: 627–642. <https://doi.org/10.1007/s00425-013-1921-1>
- Payne R., Welham S, Harding S (2013) A guide to REML in GenStat 97
- Pérez P, de Los Campos G (2014) Genome-wide regression & prediction with the BGLR statistical package. *Genetics* 198: 483–495. <https://doi.org/10.1534/genetics.114.164442>
- Piyasatian N, Fernando RL, Dekkers JCM (2006) Genomic selection for composite line development using low density marker maps. In: 8th World congress on genetics applied to livestock production. Belo Horizonte, Brazil
- Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome J* 5:92–102. <https://doi.org/10.3835/plantgenome2012.05.0005>
- Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink JL (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5:103–113. <https://doi.org/10.3835/plantgenome2012.06.0006>
- Pootakham W, Jomchai N, Ruang-areerate P, Shearman JR, Sonthirod C, Sangsrakru D, Tragoonrun S, Tangphatsomruang S (2015) Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics* 105:288–295. <https://doi.org/10.1016/j.ygeno.2015.02.002>
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 7: 574–578. <https://doi.org/10.1111/j.1471-8286.2007.01758.x>

- Rincent R, Laloë D, Nicolas S et al (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.) *Genetics* 192: 715–728
- Roberts A, McMillan L, Wang W, Parker J, Rusyn I, Threadgill D (2007) Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics* 23:401–407. <https://doi.org/10.1093/bioinformatics/btm220>
- Rooney WL, Aydin S (1999) Genetic control of a photoperiod-sensitive response in sorghum bicolor (L.) Moench. *Crop Sci* 39:397–400
- Rooney WL, Blumenthal J, Bean B, Mullet JE (2007) Designing sorghum as a dedicated bioenergy feedstock. *Biofuels Bioprod Biorefin* 1:147–157
- Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet* 123:218–223. <https://doi.org/10.1111/j.1439-0388.2006.00595.x>
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Smith A, Cullis B, Thompson R (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57:1138–1147
- Solberg TR, Sonesson A, Wooliams J, Meuwissen THE (2006) Genomic selection using different marker types and density. In: 8th World congress on genetics applied to livestock production. Belo Horizonte, Brazil
- Sonah H, Bastien M, Iqira E, Tardivel A, Légaré G, Boyle B, Normandeau É, Laroche J, Larose S, Jean M, Belzile F (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* 8:1–9. <https://doi.org/10.1371/journal.pone.0054603>
- Stacklies W, Redestig H, Scholz M, Walther D, Selbig J (2007) pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics* 23:1164–1167. <https://doi.org/10.1093/bioinformatics/btm069>
- Sticklen MB (2008) Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nat Rev Genet* 9: 433–443. <https://doi.org/10.1038/nrg2336>
- 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>
- Van Soest PJ, Robertson JB, Lewis BA (1991) Methods for dietary fiber, neutral detergent fiber, and nonstarch polysaccharides in relation to animal nutrition. *Journal of Dairy Science* 74: 3583–3597
- VanRaden PM, Van Tassell CP, Wiggans GR et al (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci* 92:16–24. <https://doi.org/10.3168/jds.2008-1514>
- Vermerris W, Saballos A, Ejeta G, Mosier NS, Ladisch MR, Carpita NC (2007) Molecular breeding to enhance ethanol production from corn and sorghum stover. *Crop Sci* 47:143–153. <https://doi.org/10.2135/cropsci2007.04.0013IPBS>
- Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L, Mastrangelo AM, Whan A, Stephen S, Barker G, Wieseke R, Plieske J, International Wheat Genome Sequencing Consortium, Lillemo M, Mather D, Appels R, Dolferus R, Brown-Guedira G, Korol A, Akhunova AR, Feuillet C, Salse J, Morgante M, Pozniak C, Luo MC, Dvorak J, Morell M, Dubcovsky J, Ganai M, Tuberosa R, Lawley C, Mikoulitch I, Cavanagh C, Edwards KJ, Hayden M, Akhunov E (2014) Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol J* 12:787–796. <https://doi.org/10.1111/pbi.12183>
- Weigel KA, de Los Campos G, Vazquez AI, Rosa GJM, Gianola D, Van Tassel CP (2010) Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. *J Dairy Sci* 93:5423–5435. <https://doi.org/10.3168/jds.2010-3149>
- Yu H, Xie W, Li J, Zhou F, Zhang Q (2013) A whole-genome SNP array (RICE6K) for genomic breeding in rice. *Plant Biotechnol J* 12:28–37. <https://doi.org/10.1111/pbi.12113>
- Yu X, Li X, Guo T, Zhu C, Wu Y, Mitchell SE, Roozeboom KL, Wang D, Wang ML, Pederson GA, Tesso TT, Schnable PS, Bernardo R, Yu J (2016) Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nat Plants* 2. <https://doi.org/10.1038/nplants.2016.150>
- Zeng Z (2001) Statistical methods for mapping quantitative trait loci. North Carolina State University, Raleigh
- Zeng ZB, Kao CH, Basten C (1999) Estimating the genetic architecture of quantitative traits. *Genet Res* 74:279–289
- 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 (2014) 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, Zeng J, Fernando R, Reif JC (2013) Genomic prediction of hybrid wheat performance. *Crop Sci* 53. <https://doi.org/10.2135/cropsci2012.08.0463>
- Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28:3326–3328. <https://doi.org/10.1093/bioinformatics/bts606>
- Zheng Y, Pan Z, Zhang R (2009) Overview of biomass pretreatment for cellulosic ethanol production. *Int J Agric Biol Eng* 2:51–68. <https://doi.org/10.3965/IJABE.V2I3.168>
- Zhong S, Dekkers JCM, Fernando RL, Jannink J-L (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. *Genetics* 182:355–364. <https://doi.org/10.1534/genetics.108.098277>