PLOS ONE

RESEARCH ARTICLE

Breeding *Jatropha curcas* by genomic selection: A pilot assessment of the accuracy of predictive models

Leonardo de Azevedo Peixoto[1]*, Bruno Galvêas Laviola[2], Alexandre Alonso Alves[2], Tatiana Barbosa Rosado[2], Leonardo Lopes Bhering[1]

1 Biology Department, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil, **2** Empresa Brasileira de Pesquisa Agropecuária, Embrapa Agroenergia, Parque Estação Biológica–PqEB s/n, Asa Norte, Brasília, Brazil

* leoazevedop@gmail.com

Citation: Azevedo Peixoto Ld, Laviola BG, Alves AA, Rosado TB, Bhering LL (2017) Breeding *Jatropha curcas* by genomic selection: A pilot assessment of the accuracy of predictive models. PLoS ONE 12(3): e0173368. https://doi.org/10.1371/journal.pone.0173368

Editor: Rongling Wu, Pennsylvania State University, UNITED STATES

Received: November 14, 2016

Accepted: February 20, 2017

Published: March 15, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: We are thankful to CAPES (Coordenacao de Aperfeicoamento de Pessoal do Ensino Superior), Cnpq (National Counsel of Technological and Scientific Development), FAPEMIG (Fundacao de Amparo a Pesquisa de Minas Gerais), Funarbe (Fundacao Arthur Bernardes) and Federal University of Vicosa for financial support. We also thank the Biometric Lab (Federal University of

Abstract

Genomic wide selection is a promising approach for improving the selection accuracy in plant breeding, particularly in species with long life cycles, such as Jatropha. Therefore, the objectives of this study were to estimate the genetic parameters for grain yield (GY) and the weight of 100 seeds (W100S) using restricted maximum likelihood (REML); to compare the performance of GWS methods to predict GY and W100S; and to estimate how many markers are needed to train the GWS model to obtain the maximum accuracy. Eight GWS models were compared in terms of predictive ability. The impact that the marker density had on the predictive ability was investigated using a varying number of markers, from 2 to 1,248. Because the genetic variance between evaluated genotypes was significant, it was possible to obtain selection gain. All of the GWS methods tested in this study can be used to predict GY and W100S in Jatropha. A training model fitted using 1,000 and 800 markers is sufficient to capture the maximum genetic variance and, consequently, maximum prediction ability of GY and W100S, respectively. This study demonstrated the applicability of genome-wide prediction to identify useful genetic sources of GY and W100S for Jatropha breeding. Further research is needed to confirm the applicability of the proposed approach to other complex traits.

Introduction

Currently, many countries have invested a lot of money into researching promising species for biofuel production [1] due to the worldwide concern over the emission of toxic gases, which enhances the greenhouse effect and contributes to climate change [2]. Thus, Jatropha (*Jatropha curcas* L.) has become a potential crop for producing biofuel due to the high oil content found in its seeds and the ability to transform this oil into biofuel [3, 4]. Jatropha has an average of 35% seed oil content, and the oil extracted from the seeds has 24.6% crude protein and 47.2% crude fat [5].

Moreover, Jatropha has several agronomic morphological traits that make it a useful crop for producing biofuel and feeding animals, such as drought tolerance [6], rapid growth, easy

Vicosa, Brazil) where all analyses were performed by remote access.

Competing interests: The authors have declared that no competing interests exist.

propagation [7], the fact that it can be grown at almost all altitudes, and because the plants can produce for more than 50 years [8]. In addition, Jatropha oil has good oxidation stability, low viscosity, and a low pour point, making its oil better than soybean oil and palm oil [9].

Although marker-assisted selection (MAS) has played an important role in plant breeding for disease and pest resistance [10, 11], its application in the improvement of quantitative traits such as grain yield (GY) and seed oil content is challenging because those traits are controlled by numerous loci with small effects. Further, the environment has a large effect on these traits, resulting in small to moderate heritability. Peixoto et al. [12] evaluated 179 half-sib families in Jatropha breeding and revealed low heritability to GY (0.35) and oil content (0.24) by REML analysis.

The complexity of traits related to GY and lack of a simple MAS approach warrant the testing of other breeding approaches. Genomic wide selection (GWS), as proposed by Meuwissen et al. [13], has become an important tool to help breeders in plant and animal breeding due to its performance as a prediction model by associating marker information with phenotypic information [13]. To be applied, GWS should have two types of population: in the training population, individual plants should be genotyped and phenotyped, and in the validation population, individual plants should just be genotyped. The primary difference between GWS and traditional forms of MAS is that in GWS, instead of using QTL mapping and a test of significant markers, all markers are included in both the training and validation populations of the GWS model and that all markers are modeled as random (i.e., not chosen for inclusion in the model based on statistical analysis). By utilizing genome-wide molecular markers, GWS is becoming a promising method for the selection of complex traits in plant breeding programs [14] and has been applied to multiple crops, including wheat, maize, and barley [15–17]. The prediction accuracies of GWS have been reported to be 28% greater than MAS and 95% as accurate as phenotypic selection for a single trait in wheat [18].

A few studies have evaluated the use of GWS in forestry breeding. Wong and Bernardo [19] first evaluated the efficiency of GWS for oil palm breeding using simulated data and demonstrated the importance of improving gain per unit time. Grattapaglia and Resende [20], using deterministic models, analyzed the use of GWS in tree genetic improvement and showed that it has great potential to accelerate breeding. This was confirmed by a simulation study of *Cryptomeria japonica* breeding [21]. The prospects of and challenges for fruit quality and disease resistance have been analyzed in apple breeding strategies using GWS [22]. A first experimental study with *Pinus taeda* demonstrated the value of GWS when the models were used at the relevant selection age in accordance with the breeding zone where marker effects were estimated [23]. Similarly, a first experimental result in eucalyptus showed that GWS is of value in understanding the quantitative trait variation in forest trees and is a powerful tool for applied tree improvement [24]. Although few studies have shown the applicability of GWS in forestry breeding, no reports can be found for Jatropha. Therefore, the objectives of this study were thus to a) estimate the genetic parameters for GY and weight of 100 seeds (W100S) using restricted maximum likelihood (REML); b) compare the performance of GWS methods to predict GY and W100S; and c) estimate how many markers are needed to train the GWS model to obtain the maximum accuracy.

Material and methods

Experiment design

Germplasm bank experiment. 179 Jatropha half-sib families from the Embrapa Cerrados germplasm bank were evaluated in this experiment. The Brazilian region where each family were collected can be found in the S1 Table. It was laid out in the experimental field of

Embrapa Cerrados, Planaltina, Distrito Federal, Brazil (15˚35'30"S and 47˚42'30"W; 1007 m asl). The experiment was implemented in November 2008 in a complete randomized block design with 2 replications and 5 plants per replication. Plants were arranged in rows, with 4 m between rows and 2 m between plants. The half-sib families were evaluated in 5 crop years from 2010 to 2014 for weight of 100 seeds (W100S) and grain yield (GY) [25]. Although the experiment was evaluated for 5 years, only the 2013 evaluation was used to perform the analysis because the diallel experiment was only evaluated in that year. All management practices were based on Dias et al. [26], and they were adapted according to recent research advances regarding Jatropha in Brazil [27, 28].

Diallel experiment. The experiment was implemented in November 2011 in a complete randomized block design with 5 replications and 3 plants per plot, with 4 m between rows and 2 m between plants. The diallel experiment was carried out using 3 segregating families, with 14 individuals per family (S2 Table). Segregating families were part of a complete diallel and were formed by crossing the contrasting genotypes of the germplasm bank, which contained genotypes of the following characteristics: nontoxic and susceptible to *Oidium* spp., toxic and resistant to *Oidium* spp., and toxic and susceptible to *Oidium* spp. (S2 Table). Trials were located at the experimental area of Embrapa Cerrados, in Planaltina-DF, Brazil (15˚35'30"S and 47˚42'30"W, 1007 m asl). Crop management practices, e.g., nutrition and pest and disease control, were carried out to maintain the germplasm bank, as recommended for the species [26, 29]. The experiment was evaluated in 2013 for GY and W100S.

Genotypic data

Because genotyping is expensive and no chip has been established for Jatropha, only 78 plants were genotyped. Thirty-six plants (the first plant in block one) were from the germplasm bank experiment, and 42 plants (14 plants per crossing) were from the diallel experiment (S3 Table).

Total genomic DNA was extracted from younger leaves using the protocol of [30] with minor modifications. Briefly, 5 g of leaves was ground to a powder in liquid nitrogen, 20 μl of extraction buffer (2% CTAB, 20 mM EDTA, 2% PVP, 1.4 M NaCl, 100 mMTris–HClpH 8.0 and 1% β-mercaptoethanol) was added, and the homogenized samples were incubated at 65˚C for 1 h. The supernatant was extracted twice with chloroform:isoamyl alcohol (24:1, v/v) and treated with RNase A (100 mg/ml) at 37˚C for 30 min. DNA was precipitated with isopropanol and washed twice with 70% ethanol. Pelleted DNA was air dried, resuspended in 100 μl of sterile ultra-pure water, and stored at -20˚C. DNA concentration was measured using a NanoDrop spectrophotometer (NanoDrop Products, Wilmington, DE, USA), and the concentration of each sample was adjusted to 2–5 ng.μ^{-1}.

Diversity Arrays techonology (DArT PL) was the company responsible to obtain DarTs and SNPs. Many methods have been developed to reduce genome complexity, however the DArT methods provide a significant advantage via an intelligent selection of genome fraction corresponding predominantly to active genes. This selection is achieved through the use of a combination of Restriction Enzymes which separate low copy sequences (most informative for marker discovery and typing) from the repetitive fraction of the genome. While the initial DArT implementation on the microarray platform involves fluorescent labeling of representations and hybridization to dedicated DArT arrays, the DArTseq method deploys sequencing of the representations on the Next Generation Sequencing (NGS) platforms. The advantage of DArTseq over the array version of DArT is currently limited to applications requiring very high marker densities (tens of thousands of markers). This technology is therefore positioned in the area of high resolution mapping and detailed genetic dissection of traits. As modern breeding, moves rapidly in this direction, especially in larger organizations, DArTseq is

increasingly used in crop improvement applications. DArTseq for a new organism starts with optimization of complexity reduction method(s). While the choice of restriction enzyme combinations is large, DArT PL has invested considerable effort in testing various combinations on a significant number of organisms and has developed sets of complexity reduction methods (representations) that are performing quite well compared to other methods. The optimization process usually selects one dominant method of complexity reduction for the crop, but in many cases several methods were identified which offer application-specific advantages. The difference between the methods can be both quantitative (different number of unique fragments in the representation) as well as compositional (different sets of fragments captured in the representations). These differences in representation size and composition translate to different efficiencies in marker detection rate and quality (call rate and reproducibility) and can be further optimized for performance in different applications. It used 1,248 SNPs and DArTs markers.

Statistical analysis

Genetic and environmental parameters were estimated by restricted maximum likelihood (REML) analysis using the Selegen software [31]. Experiment was evaluated using the follow model:

$$Y = X_r + Z_a + W_p + e$$

where Y is the phenotypic values vector; r is the block effect vector (fixed effect); a is the additive effect vector (random effect); p is the interaction between the block effect and the genotype effect (random effect); e is the residual effect vector (random effect); and X, Z and W are the incidence matrix to the block effect, the additive effect, and the interaction between the block effect and the genotype effect, respectively.

Because the second experiment was a diallel, the dominance effect should be fitted into the mixed model. Therefore, experiment 2 was analyzed using the follow model:

$$Y = X_r + Z_a + W_p + T_f + e$$

where Y is the phenotypic values vector; r is the block effect vector (fixed effect); a is the additive effect vector (random effect); p is the interaction between the block effect and the genotype effect (random effect); f is the dominance effect vector (random effect); e is the residual effect vector (random effect); and X, Z, W and T are the incidence matrix to the block effect, the additive effect, the interaction between the block effect and the additive effect, and the dominance effect, respectively.

Genetic diversity was estimated by multidimensional scaling analysis (MDS) [32] using the MASS package in the R software [33]. The distance matrix was estimated based on markers which were identical by state (IBS), and a two-dimensional graphic was plotted based on the distance matrix using the scatterplot3d package in R.

Narrow sense heritability was calculated for each experiment as the additive genetic variance divided by the total phenotypic variance. The genetic variance was calculated using the equation proposed by Falconer et al. [34].

Prediction ability was assessed as the Pearson Correlation of the genomic estimate breeding value (GEBV) and phenotypic value in the validation population.

Genomic prediction models

Eight GWS methods were used for analysis in the field experiment: RR-BLUP, G-BLUP, Bayesian Ridge Regression (BRR), Bayes A, Bayes B, Bayes Cπ, Bayesian LASSO (BLASSO)

and Reproducing Kernel Hilbert Spaces Regression (RKHS). In all models, the phenotypic records were described as

$$y_i = \mu + g_i + \varepsilon_i$$

where $y_i = n_i^{-1}\sum_{i=1}^{k} y_{ik}$ is the average performance of the i_{th} line; n_i is the number of replicates used for computing the mean value of the i_{th} genotype; μ is an intercept; g_i is the genetic value of the i_{th} genotype; and ε_i is a model residual. The genomic selection models differed in how molecular marker information was included in g_i.

Three methods used in this work were described by Meuwissen et al. [35]: RR-BLUP, Bayes A and Bayes B. RR-BLUP assumes that each marker had variance equal to V_G/M, where V_G is the genetic variance and M is the number of markers. In the Bayes A method, each effect i is drawn from a normal distribution with its own variance: $N(0, \sigma^2_{gi})$; the variance parameters are in turn sampled from a scaled inverted chi-squared distribution. In the Bayes B approach, the prior for the proportion of markers associated with zero phenotypic variance, π, was assumed to be unknown. The other prior hyperparameters for marker variance components in Bayes A and Bayes B were as given by Meuwissen et al. [35].

G-BLUP assumes an equal variance for each marker and uses a genomic relationships matrix among all individuals in a reference set and a test set that allows it to compute the variance components and best linear unbiased predictions (BLUP) from a mixed model [36]. This was achieved by replacing the pedigree-based relationship matrix with the genomic relationship matrix (G) estimated from SNP marker genotypes to define the covariance among breeding values.

BRR assumes that each marker had a variance equal to V_G/M, where V_G is the genetic variance and M is the number of markers. The variance was assigned an inverse chi-square ($\sigma^2 \sim \chi^{-2}(S,v)$).

Bayes Cπ assumes common marker variances and allows some markers to have no effect [37]. Additionally, Bayes Cπ jointly estimates π from the training data to avoid an incorrect π that can negatively affect prediction accuracy [38].

In the BLASSO method, marker effects are assigned independent Gaussian priors with marker-specific variances ($\sigma^2_e\tau^2_j$). At the next level of the hierarchical model, the τ^2_js are assigned iid exponential priors $EXP\,[\tau^2_j|\lambda^2]$. At a deeper level of the hierarchy, λ^2 is assigned a Gamma prior with a rate (δ) and shape (r), which, in this study, were the default in the BGLR package in R. Finally, independent scaled inverse chi-square priors were assigned to the variance parameters, and the scale and degree of freedom parameters were set to $S_u = S_e = 1$ and d: $f_{.e} = d{:}f_{.u} = 4$, respectively. BLASSO is described by De Los Campos et al. [39].

In RKHS, genetic values were viewed as a Gaussian process. When markers and a pedigree were available, genetic values were modeled as the sum of two components:

$$g_i = u_i + f_i$$

where u_i is the mean and f_i is a Gaussian process with a (co)variance function proportional to the evaluations of a reproducing kernel, $K(x_i, x_j)$, evaluated in marker genotypes; here, x_i and x_j are vectors of marker genotype codes for the i_{th} and j_{th} individuals, respectively. All hyper parameters were assumed following De Los Campos et al. [40].

Marker density

The effect of numbers of markers on prediction ability was determined through five-fold cross-validation by excluding, after each interaction, the marker that had the smallest effect.

Therefore, the number of markers decreased from 1248 to 2. Each interaction was repeated 50 times to avoid sampling bias for markers, and the average of these replications was used to represent the prediction ability of each interaction.

The prediction ability average with standard error as the error bars was plotted versus the number of markers in each interaction using Boxplot. G-BLUP was used to perform these analyses because it was the fastest GWS method.

Software and computer information

All statistical modeling was performed in R. RR-BLUP and G-BLUP were performed using the rrBLUP package (function mixed.solve and kin.BLUP, respectively). The Bayes A, Bayes B, Bayes Cπ and RKHS models were performed using the BGLR package (function BGLR), and BLASSO and BRR were performed using the BLR package (function BLR).

A total of 20,000 burn-ins (number of iterations before the Bayesian analysis convergence) and 40,000 saved iterations, as obtained from the Markov chain Monte Carlo (MCMC) method, was used in all Bayesian methods. The convergence of Bayesian models was checked by inspecting trace plots of the variance parameters.

Two high-performance computers (12^{th} generation, Intel Xeon E5-26 processor, 3.30 GHz, 64 or 96 GB RAM, 1024 GB hard drive) were used to perform all analyses.

Results

Phenotypic analysis

Genetic variance was similar in both experiments for grain yield (GY), and 3 times greater in experiment 1 for weight of 100 seeds (W100S) (Table 1). The heritability was moderate for GY in experiment 1 and overestimated in experiment 2. Conversely, the heritability for W100S was overestimated in experiment 1 and moderate in experiment 2. CV_e was high and low for GY and W100S, respectively. CV_r was greater than 1 for W100S in experiment 1, but it was lower than 1 for GY in the same experiment.

Diversity analysis

Cluster analysis by MDS showed the diversity between Jatropha genotypes, and only one group was detected (Fig 1). This group was composed of the three full-sib families from the

Table 1. Genetic and environmental parameters estimated by REML analysis.

Parameters	Experiment 1		Experiment 2	
	GY	W100S	GY	W100S
σ_a^2	96,296.94	57.34	104162.66	18.85
σ_f^2	-	-	19279.05	26.17
σ_b^2	195,813.99	5.06	6192.58	0.33
σ_p^2	360420.77	30.11	108908.36	54.17
h_a^2	0.27	1.90	0.96	0.35
CV_g	13.16	5.44	-	-
CV_e	40.76	3.86	-	-
CV_r	0.33	1.40	-	-

GY–Grain Yield; W100S –Weight of 100 seeds; σ_a^2–additive variance; σ_f^2–family variance (diallel experiment); σ_b^2–variance between plots; σ_p^2–phenotypic variance; h_a^2–additive heritability; CV_g–coefficient of variation genetic; CV_e–coefficient of variation residual; and CV_r–ratio between CV_g and CV_e.

https://doi.org/10.1371/journal.pone.0173368.t001

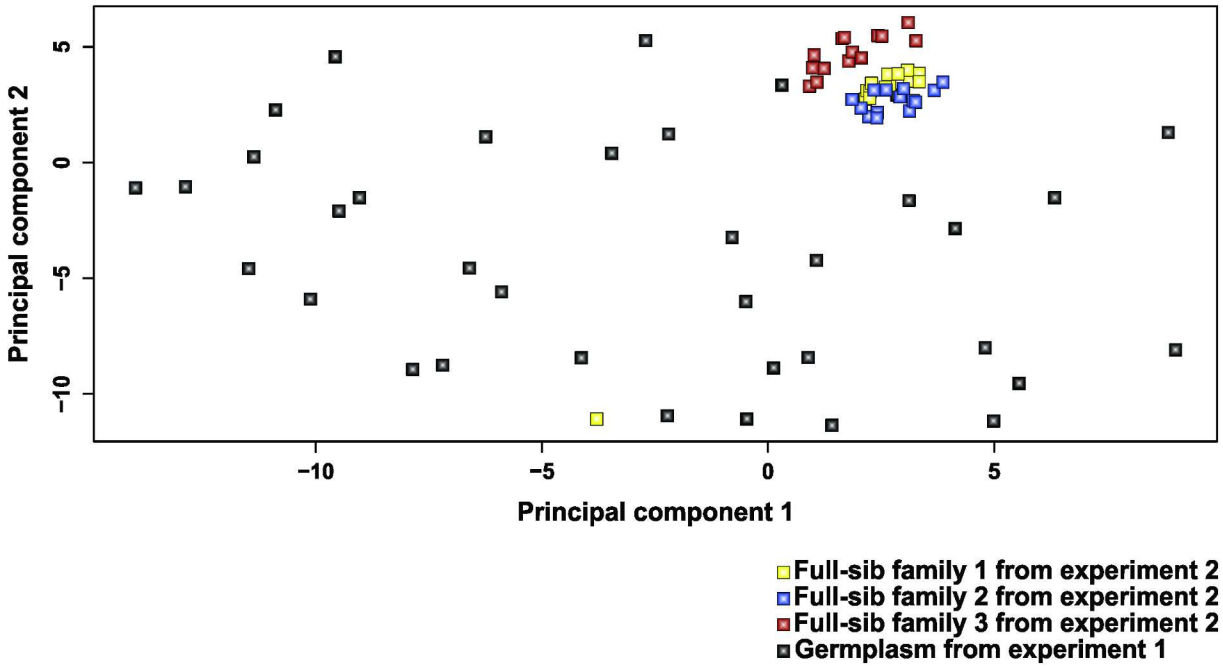

Fig 1. Multidimensional scaling analysis (MDS) showing the first two principal components based on 1,248 markers that were run on the 78 genotypes of Jatropha.

https://doi.org/10.1371/journal.pone.0173368.g001

diallel experiment, and all of the genotypes were in this group, except for one from full-sib family 1. In contrast, several genotypes are spread out in the graphic, showing the variability between the genotypes studied.

Comparison between genomic selection methods

Five-fold cross validation was performed using the full set of 1,248 markers to predict GY and W100S in Jatropha. Prediction ability was estimated as the correlation of GEBV and phenotype values in the validation population.

The prediction ability was similar between GWS methods for GY and W100S (Fig 2), except BLASSO, which presented smaller values for GY. The average prediction ability of GY (0.66) was higher than W100S (0.46).

There were no differences between the GWS methods to estimate heritability for both traits, except in that Bayes Cπ estimated a higher heritability (Fig 2).

The processing time ranged from 0.08 and 0.07 (RR-BLUP) to 753.09 and 684.23 seconds (BL) for GY and W100S, respectively. We observed that RR-BLUP and G-BLUP were the fastest methods, followed by RKHS: G-BLUP and RR-BLUP were, respectively, 500 and 100 times faster than the fastest Bayesian method (RKHS) and 7000 and 1400 times faster than the slowest Bayesian method (BLASSO).

Influence of marker density on GWS models prediction

The number of markers did not affect the prediction accuracies that presented values close to 1, except when the number of markers was less than 50 or greater than 1,000 (Fig 3A and 3B) for both traits. The estimated heritability showed the same shape; i.e., heritability increased beyond 50 markers, stayed constant until 1,000 and 800 markers for GY and W100S, respectively, and decreased when more than 1,000 and 800 markers were used to train the model (Fig 3A and 3B).

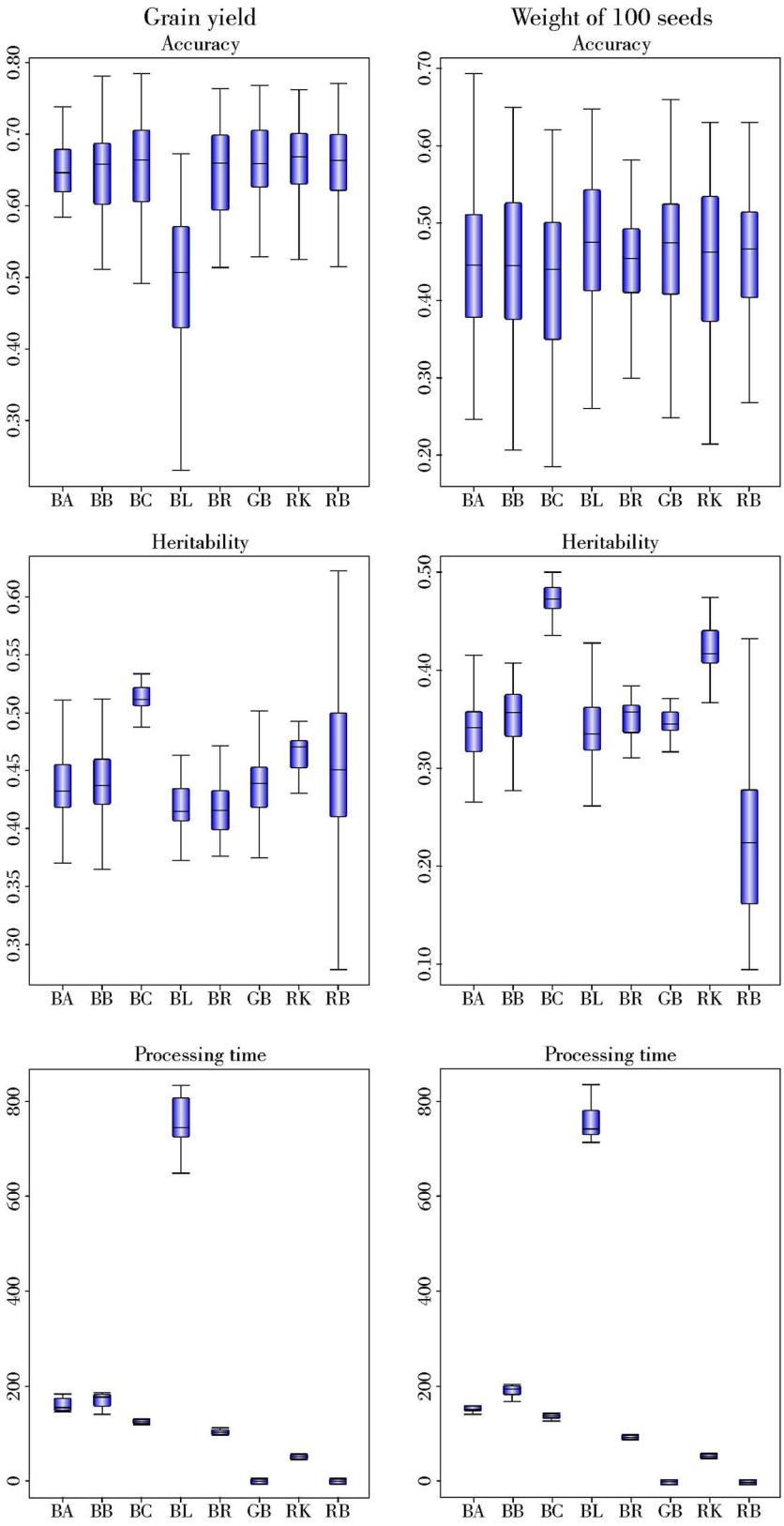

Fig 2. Comparison between genomic selection methods to predict grain yield and weight of 100 seeds. BA- Bayes A; BB–Bayes B; BC–Bayes Cπ; BR–Bayesian Ridge Regression; BL–Bayesian LASSO; GB–G-BLUP; RK–Reproducing kernel Hilbert Space; and RB–RR-BLUP.

https://doi.org/10.1371/journal.pone.0173368.g002

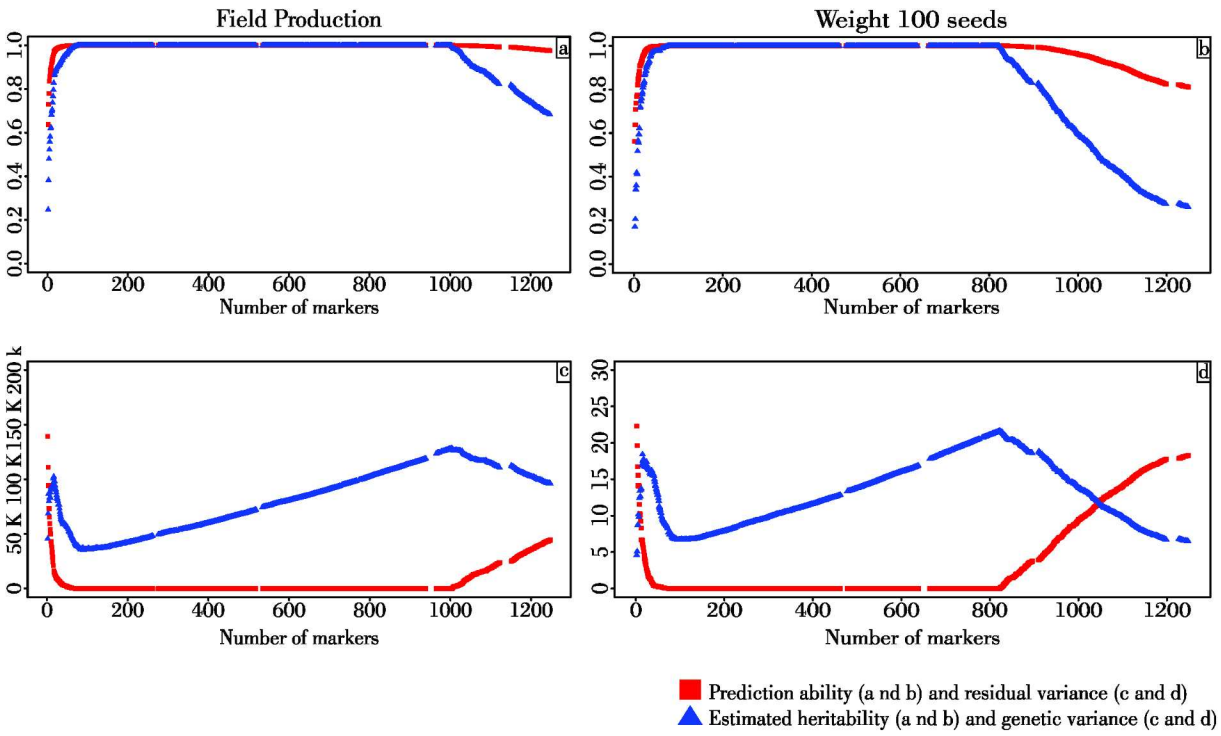

Fig 3. Effect of the number of markers in the prediction ability and estimated heritability (a and b); and genetic and residual variance (c and d) of Grain Yield (GY) and Weight of 100 Seeds (W100S).

https://doi.org/10.1371/journal.pone.0173368.g003

Genetic variance presented a cubic shape, decreasing beyond 100 markers, increasing beyond 1,000 and 800 for GY and W100S, respectively, and decreasing when more than 1,000 and 800 markers were used to train the model (Fig 3C and 3D). The residual variance decreased over 100 markers, coming to values near 0, maintained small values beyond 1,000 and 800 markers for GY and W100S, respectively, and increased thereafter (Fig 3C and 3D).

Discussion

Phenotypic analysis

The genetic variance found in our study for GY was similar to those of previous reports [12, 25] while the heritability was greater. The heritability was overestimated in our study because the residual variance was overestimated because the residual variance was calculated as the difference between the phenotypic variance and other components of variance. Then, when the genetic variance is too large and the number of individuals is small, the residual variance can be negative and then the heritability can be greater than 1. This problem can be solved evaluating more plants. However, we could not do it because we had problems with the DNA extraction, and just a few plants could be evaluated. When we estimated these parameters using all plants in the experiment this problem was solved as showed by Junqueira et al. [25] and Peixoto et al. [41].

It was observed that the genetic variance for GY was greater than that for W100S. This difference was possibly due to a scale effect: whereas GY ranged from 0 to 999, W100S ranged from 0 to 96.78.

Another parameter commonly used to evaluate the genetic variability between families is the ratio coefficient of variation (CV_r) [42]. When the relationship between CV_g and CV_e is

greater than 1, the selection gain will be high [43]. In this study, CV_r was greater than 1 for W100S in experiment 1. However, CV_r was lower than 1 for GY, which indicated that phenotypic selection may not provide any genetic gain for this trait.

Based on the parameters estimated by REML analysis, selecting superior genotypes for GY and W100S based on phenotypic values would not provide a good selection gain for the next generation because approximately 73% and 65% of the phenotypic variance is not genetic for GY and W100S, respectively. Thus, it is necessary to use more accurate methodologies to predict genetic effects. Therefore, based on these results and on previous research [12, 25], we suggest that the GWS is more appropriate than ANOVA and REML/BLUP to perform analyses and select superior genotypes for Jatropha breeding because the GWS can capture minor genetic differences between families, whereas ANOVA and REML/BLUP cannot.

Comparison between GWS methods

Several studies have shown that compared with ridge regression methods, more complex statistical methods give only a small increase in the accuracy of genomic prediction for polygenic traits [44–47]. However, this small increase in the prediction accuracy is not sufficient to make Bayesian methods generate statistically better results than G-BLUP. GWS studies conducted in maize, wheat, oat, and barley for both agronomic and disease traits also suggested slight differences among various genomic prediction algorithms [15, 48–50]. In this study, G-BLUP performed similar to all Bayesian methods for GY and W100S. This might be due to the use of non-informative prior distributions in Bayesian methods, resulting in the posterior distribution being influenced solely by the likelihood function. Perhaps meta-analysis can improve accuracies in Bayesian methods by fitting prior distributions using parameters estimated by historical data [51]. Moreover, G-BLUP has other advantages, such as relative simplicity, reduced computing time, and the well-known optimality properties of mixed models for selection [52]. For example, Azevedo et al. [53] analyzed 10 GWS models, including G-BLUP and BLASSO, proposed modifications to the models, and concluded that the G-BLUP, BAYES A^*B^* (−2,8) and BAYES A^*B^* (4,6) methods presented the best results and were adequate for accurately predicting genomic breeding. Thus, G-BLUP was chosen to perform other analyses.

Moreover, it has now been demonstrated that predictive models built on the basis of genome-wide markers allow breeders to obtain higher selective accuracy, even for traits of low heritability, such as GY and seed oil content in Jatropha. In addition, genomic breeding values may be estimated at the seedling stage, which can reduce the Jatropha breeding cycle by at least five years (6 years for breeding cycles with GWS versus 12 years for breeding cycles without GWS). Because selection response is inversely proportional to breeding cycle length, we calculated the expected impact of GWS on Jatropha breeding. Considering the accuracies and cycles reported here, GWS may increase the selection efficiency in Jatropha breeding by more than 100% [20, 24, 54]. Technow et al. [55], to show how GWS can override phenotypic selection, proposed a formula to estimate the response to indirect selection, being: $L_Y < (r_A/H_X)\, L_x$; where L_Y is the cycle length of GWS, r_A is the genomic prediction accuracy, H_X is the phenotypic selection accuracy, and L_X is the cycle length of phenotypic selection. Substituting the values estimated for GY and W100S in the formula, it can be observed that GWS must be superior to phenotypic selection if the cycle length of GWS is the same cycle length of phenotypic selection for GY and less than 49% the cycle length of phenotypic selection for W100S. Therefore, because the cycle is half the length when using GWS (6 years = 1 year for crossing and 5 years to evaluate in different environments) instead of a traditional breeding cycle (12 years = 7 years for crossing and 5 years to evaluate in different environments), GWS is a useful tool to reduce the breeding cycle.

In the case of perennial crops such as Jatropha, they need several years, ranging from 10 to 14 years, to obtain suitable phenotypic evaluations. Based on practical considerations and the theoretical equation presented above, GWS may improve the efficiency of breeding programs. The main step in which GWS will be useful is shortening the length of the breeding cycle. This will occur because the progeny testing phase can be omitted when GWS is applied, and breeders will thus be able to perform early selection at the seedling stage. Then, selected individuals can be immediately propagated by micropropagation protocols; consequently, optimized clonal trials with several years of anticipation can be established compared to a classical breeding.

The selection response per time unit may be drastically increased (by as much as 50%) when the breeding cycle is reduced because the selection response is inversely proportional to the breeding cycle length, as theoretically and experimentally demonstrated [20]. For instance, simulation studies for oil palm have demonstrated that GWS can be more effective than phenotype selection in terms of both cost and time reduction because breeders can perform four breeding cycles in the same period of time when using GWS instead of the two breeding cycles that are permitted when traditional breeding is used [19, 24]. Moreover, with the development of genotyping-by-sequencing approaches, early selection may also allow breeders to increase selection intensity, thus allowing them to have a large number of individuals quickly genotyped for thousands of markers at a low cost. Additionally, experiments in forest breeding are usually limited in size due to economic and operational aspects, which reduces both the number of evaluated individuals and the accuracy of phenotypic selection. Therefore, breeders will be able to reduce their investment in field-testing using GWS by evaluating just a few individuals that will be used to train the model, thereby saving time and resources and improving the selection precision for traits of low heritability.

Influence of marker density on GWS models prediction

The effectiveness of GWS depends on the correlation between the predicted genotypic value and the underlying true genotypic value [56]. This correlation, also called the prediction ability, of GWS has been expressed as a function of the marker density, training population size (N), trait heritability on an entry-mean basis (h^2), and the effective number of quantitative trait loci (QTL) or effective number of chromosome segments underlying the trait (Me) [57, 58]. Simulation and cross-validation studies have indicated that prediction accuracy generally increases as h^2 increases [18, 23, 49] and is not affected when the number of markers increases [59, 60]. Peixoto et al. [12] showed, using REML/BLUP, that the most important traits in Jatropha have different heritabilities, such as GY, oil content, phorbol ester concentration, and W100S, the heritabilities of which were 0.32, 0.24, 0.71 and 0.85, respectively. Therefore, different strategies should be developed to use GWS in Jatropha and obtained high prediction abilities for those traits.

Models fitted using over 1000 and 800 markers were capable of predicting GY and W100S, respectively (Fig 3). The comparable performance of a limited number of markers (1000 and 800) relative to the complete marker data set could be due to marker saturation because random markers with uniform coverage across each chromosome were selected. With a larger linkage disequilibrium, the addition of more markers will not increase the accuracy of the predictive models [61]. Because a linear correlation between the number of markers and prediction accuracy was not observed in this research, a good GWS model for predicting GY and W100S in Jatropha can be fitted by using approximately 1000 and 800 markers, respectively, in a diverse genotype collection. The use of a small SNP set can lead to cost savings. Using a uniform or common SNP set will allow the consistent use of genome-wide prediction in research and breeding programs.

However, before Jatropha breeding programs incorporate GWS on a large scale, the results found in our study must be validated across years and by evaluating progenies. Because the genetic material used in this study consisted of diverse accessions from the germplasm bank (Fig 1), the results from the genetic structure and composition of entries in this study would be applicable to germplasm enhancement programs using diverse collections to obtain parental materials.

Future GWS applications in Jatropha

Because the demand for biodiesel is constantly increasing, the development of dedicated crops has been suggested as a strategic action. Thus, biodiesel production is expected to become much more efficient if not only conversion processes themselves are improved but also oil feedstocks are optimized to this end. In that context, genomics offers innumerable technologies for collecting genetic information that could be potentially integrated into Jatropha breeding to aid in the development of cultivars with outstanding performance for biodiesel production. Because genomics offers a platform to learn more about the relationships of genes and phenotypes, the long-term goal of applying genomics to breeding is to link genomic information with the field research that is currently underway, with the purpose of developing accurate predictive models. Such models could then be operationally used by breeders to estimate the performance and adaptability of genotypes across locations or ecosystems based on genetic data alone, i.e., without the need for conducting laborious and expensive phenotyping trials at the beginning of the breeding cycle. In the context of a long-lived perennial crop, with long breeding cycles and late-expressing traits, the achievement of such a long-term goal promises to revolutionize selective breeding [62]. Because some of the most promising feedstocks for biodiesel production, such as Jatropha, oil palm, macaw palm (*Acrocomia aculeate*), and pongamia (*Pongamia pinnata*), are perennial crops, genomic breeding is one of the most promising ways to foster the development of perennial crops dedicated to biodiesel production.

In the near future, GWS can improve the efficiency of producing Jatropha oil, but many studies are needed to prove all such theories in practice experiments because no study to date has evaluated GWS in biofuel traits. Therefore, studies should evaluate how the GWS method is better to capture high accuracy for oil production, how many individuals and markers are needed to train the model, and how the GxE interaction can influence prediction accuracy.

Conclusion

There was genetic variance between the genotypes evaluated, and it was possible to obtain selection gain using GWS;

All genomic selection methods tested in this study can be used to predict the grain yield and weight of 100 seeds in Jatropha;

Training models fitted using 1,000 and 800 markers are sufficient to capture the maximum genetic variance and, consequently, the maximum prediction ability for grain yield and weight of 100 seeds, respectively.

Supporting information

S1 Table. Identification and origin for each accession used in the germplasm bank experiment.
(DOCX)

S2 Table. Identification of families used in the diallel experiment.
(DOCX)

S3 Table. Data set used to perform all genomic wide selection analyses in this study.
(DOCX)

Author Contributions

Conceptualization: LAP.

Data curation: BGL AAA TBR.

Formal analysis: LAP LLB.

Funding acquisition: BGL AAA.

Investigation: BGL AAA.

Methodology: LAP LLB.

Project administration: BGL LLB.

Resources: BGL AAA TBR.

Software: LAP.

Supervision: BGL LLB.

Validation: AAA.

Visualization: LAP.

Writing – original draft: LAP.

Writing – review & editing: BGL LLB.

References

1. Bailis R, Baka J. Constructing sustainable biofuels: governance of the emerging biofuel economy. Annals of the Association of American Geographers. 2011; 101(4):827–38.
2. Naylor RL, Liska AJ, Burke MB, Falcon WP, Gaskell JC, Rozelle SD, et al. The ripple effect: biofuels, food security, and the environment. Environment: Science and Policy for Sustainable Development. 2007; 49(9):30–43.
3. Berchmans HJ, Hirata S. Biodiesel production from crude *Jatropha curcas* L. seed oil with a high content of free fatty acids. Bioresource technology. 2008; 99(6):1716–21. https://doi.org/10.1016/j.biortech.2007.03.051 PMID: 17531473
4. Pu Y, Treasure T, Gonzalez RW, Venditti R, Jameel H. Autohydrolysis pretreatment of mixed hardwoods to extract value prior to combustion. BioResources. 2011; 6(4):4856–70.
5. Akintayo ET. Characteristics and composition of *Parkia biglobbossa* and *Jatropha curcas* oils and cakes. Bioresource technology. 2004; 92(3):307–10. https://doi.org/10.1016/S0960-8524(03)00197-4 PMID: 14766165
6. Becker K, Makkar HPS. *Jatropha curcas*: a potential source for tomorrow's oil and biodiesel. Lipid Technology. 2008; 20(5):104–7.
7. Openshaw K. A review of *Jatropha curcas*: an oil plant of unfulfilled promise. Biomass and Bioenergy. 2000; 19(1):1–15.
8. Nithiyanantham S, Siddhuraju P, Francis G. Potential of *Jatropha curcas* as a biofuel, animal feed and health products. Journal of the American Oil Chemists' Society. 2012; 89(6):961–72.
9. Bailis R, McCarthy H. Carbon impacts of direct land use change in semiarid woodlands converted to biofuel plantations in India and Brazil. GCB Bioenergy. 2011; 3(6):449–60.
10. Butler JB, Freeman JS, Vaillancourt RE, Potts BM, Glen M, Lee DJ, et al. Evidence for different QTL underlying the immune and hypersensitive responses of *Eucalyptus globulus* to the rust pathogen Puccinia psidii. Tree Genetics & Genomes. 2016; 12(3):1–13.

11. Terakami S, Moriya S, Adachi Y, Kunihisa M, Nishitani C, Saito T, et al. Fine mapping of the gene for susceptibility to black spot disease in Japanese pear (Pyrus pyrifolia Nakai). Breeding science. 2016; 66(2):271. https://doi.org/10.1270/jsbbs.66.271 PMID: 27162498
12. Peixoto LA, Laviola BG, Bhering LL, Mendonça S, Costa TdSA, Antoniassi R. Oil content increase and toxicity reduction in jatropha seeds through family selection. Industrial Crops and Products. 2016; 80:70–6.
13. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157(4):1819–29. PMID: 11290733
14. Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, et al. 2 Genomic Selection in Plant Breeding: Knowledge and Prospects. Advances in agronomy. 2011; 110:77.
15. Lorenz AJ, Smith KP, Jannink J-L. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop science. 2012; 52(4):1609–21.
16. Burgueño J, de los Campos G, Weigel K, Crossa J. Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Science. 2012; 52(2):707–19.
17. Heslot N, Yang H-P, Sorrells ME, Jannink J-L. Genomic selection in plant breeding: a comparison of models. Crop Science. 2012; 52(1):146–60.
18. Heffner EL, Jannink J-L, Iwata H, Souza E, Sorrells ME. Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Science. 2011; 51(6):2597–606.
19. Wong CK, Bernardo R. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theoretical and Applied Genetics. 2008; 116(6):815–24. https://doi.org/10.1007/s00122-008-0715-5 PMID: 18219476
20. Grattapaglia D, Resende MDV. Genomic selection in forest tree breeding. Tree Genetics & Genomes. 2011; 7(2):241–55.
21. Iwata H, Hayashi T, Tsumura Y. Prospects for genomic selection in conifer breeding: a simulation study of *Cryptomeria japonica*. Tree genetics & genomes. 2011; 7(4):747–58.
22. Kumar S, Bink MC, Volz RK, Bus VG, Chagné D. Towards genomic selection in apple (*Malus*× *domestica* Borkh.) breeding programmes: prospects, challenges and strategies. Tree Genetics & Genomes. 2012; 8(1):1–14.
23. Resende MF, Muñoz P, Resende MD, Garrick DJ, Fernando RL, Davis JM, et al. Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics. 2012; 190(4):1503–10. https://doi.org/10.1534/genetics.111.137026 PMID: 22271763
24. Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, et al. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytologist. 2012; 194(1):116–28. https://doi.org/10.1111/j.1469-8137.2011.04038.x PMID: 22309312
25. Junqueira VS, Peixoto LdA, Laviola BG, Bhering LL, Mendonça S, Costa TdSA, et al. Bayesian Multi-Trait Analysis Reveals a Useful Tool to Increase Oil Concentration and to Decrease Toxicity in *Jatropha curcas* L. PloS one. 2016; 11(6):e0157038. https://doi.org/10.1371/journal.pone.0157038 PMID: 27281340
26. Dias LAdS, Leme LP, Laviola BG, Pallini Filho A, Pereira OL, Carvalho M, et al. Cultivo de pinhão-manso (*Jatropha curcas* L.) para produção de óleo combustível. Viçosa, MG. 2007; 1:1–40.
27. Carels N, Sujatha M, Bahadur B. Jatropha, Challenges for a New Energy Crop: Farming, Economics and Biofuel. Verlag: Springer 2013.
28. Bahadur B, Sujatha M, Carels Np. Jatropha, Challenges for a New Energy Crop: Genetic Improvement and Biotechnology: Springer Science & Business Media; 2012.
29. Laviola BG, Rosado TB, Bhering LL, Kobayashi AK, Resende MDVd. Genetic parameters and variability in physic nut accessions during early developmental stages. Pesquisa Agropecuária Brasileira. 2010; 45(10):1117–23.
30. Grattapaglia D, Sederoff R. Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. Genetics. 1994; 137(4):1121–37. PMID: 7982566
31. Resende MDVd. Software SELEGEN—REML/BLUP. Colombo: EMBRAPA Floresta; 2002.
32. Mohammadi SA, Prasanna BM. Analysis of genetic diversity in crop plants—salient statistical tools and considerations. Crop Science. 2003; 43(4):1235–48.
33. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.

34. Falconer DS, Mackay TF, Frankham R. Introduction to Quantitative Genetics (4th edn). Trends in Genetics. 1996; 12(7):280.

35. Meuwissen THT, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157(4):1819–29. PMID: 11290733

36. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science. 2009; 92(2):433–43. https://doi.org/10.3168/jds.2008-1646 PMID: 19164653

37. Jannink JL, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics. 2010; 9(2):166–77. https://doi.org/10.1093/bfgp/elq001 PMID: 20156985

38. Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R. Additive genetic variability and the Bayesian alphabet. Genetics. 2009; 183(1):347–63. https://doi.org/10.1534/genetics.109.103952 PMID: 19620397

39. De Los Campos G, Naya H, Gianola D, Crossa J, Legarra A, Manfredi E, et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 2009; 182(1):375–85. https://doi.org/10.1534/genetics.109.101501 PMID: 19293140

40. De Los Campos G, Gianola D, Rosa GJM, Weigel KA, Crossa J. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research. 2010; 92(04):295–308.

41. Peixoto L, Bhering L, Cruz C. Determination of the optimal number of markers and individuals in a training population necessary for maximum prediction accuracy in F2 populations by using genomic selection models. Genetics and molecular research: GMR. 2016; 15(4).

42. Resende MDV. Análise estatística de modelos mistos via REML/BLUP na experimentação em melhoramento de plantas perenes. Colombo: Embrapa Florestas; 2000.

43. Vencovsky R. Herança quantitativa. In: Paterniani E, editor. Melhoramento e a produção de milho no Brasil. Piracicaba: Fundação Cargill; 1987. p. 137–214.

44. Bernardo R, Yu J. Prospects for genomewide selection for quantitative traits in maize. Crop Science. 2007; 47(3):1082–90.

45. Daetwyler HD, Calus MP, Pong-Wong R, de los Campos G, Hickey JM. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics. 2013; 193(2):347–65. https://doi.org/10.1534/genetics.112.147983 PMID: 23222650

46. Cleveland MA, Hickey JM, Forni S. A common dataset for genomic analysis of livestock populations. G3: Genes| Genomes| Genetics. 2012; 2(4):429–35.

47. Clark SA, Hickey JM, Daetwyler HD, van der Werf JH. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol. 2012; 44(4).

48. Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink J-L. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. The Plant Genome. 2011; 4 (2):132–44.

49. Lorenzana RE, Bernardo R. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theoretical and applied genetics. 2009; 120(1):151–61. https://doi.org/10.1007/s00122-009-1166-3 PMID: 19841887

50. Rutkoski J, Benson J, Jia Y, Brown-Guedira G, Jannink J-L, Sorrells M. Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. The Plant Genome. 2012; 5(2):51–61.

51. Akanno E, Schenkel F, Quinton V, Friendship R, Robinson J. Meta-analysis of genetic parameter estimates for reproduction, growth and carcass traits of pigs in the tropics. Livestock Science. 2013; 152 (2):101–13.

52. Fernando R, Gianola D. Optimal properties of the conditional mean as a selection criterion. Theoretical and applied genetics. 1986; 72(6):822–5. https://doi.org/10.1007/BF00266552 PMID: 24248207

53. Azevedo CF, de Resende MDV, e Silva FF, Viana JMS, Valente MSF, Resende MFR, et al. Ridge, Lasso and Bayesian additive-dominance genomic models. BMC genetics. 2015; 16(1):1.

54. Resende MFR, Munoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, et al. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytologist. 2012; 193(3):617–24. https://doi.org/10.1111/j.1469-8137.2011.03895.x PMID: 21973055

55. Technow F, Bürger A, Melchinger AE. Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. G3: Genes| Genomes| Genetics. 2013; 3(2):197–203. https://doi.org/10.1534/g3.112.004630 PMID: 23390596

56. Goddard ME, Hayes BJ. Genomic selection. Journal of Animal breeding and Genetics. 2007; 124 (6):323–30. https://doi.org/10.1111/j.1439-0388.2007.00702.x PMID: 18076469

57. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010; 185(3):1021–31. https://doi.org/10.1534/genetics.110.116855 PMID: 20407128

58. Daetwyler HD, Villanueva B, Woolliams JA. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS one. 2008; 3(10):e3395. https://doi.org/10.1371/journal.pone.0003395 PMID: 18852893

59. Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, et al. Correction: 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. 2015; 11(6):e1005350. https://doi.org/10.1371/journal.pgen.1005350 PMID: 26125618

60. Pszczola M, Strabel T, Mulder HA, Calus MPL. Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of dairy science. 2012; 95(1):389–400. https://doi.org/10.3168/jds.2011-4338 PMID: 22192218

61. Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redoña E, et al. 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. 2015; 11(2):e1004982. https://doi.org/10.1371/journal.pgen.1004982 PMID: 25689273

62. Neale DB, Kremer A. Forest tree genomics: growing resources and applications. Nature Reviews Genetics. 2011; 12(2):111–22. https://doi.org/10.1038/nrg2931 PMID: 21245829