2017 International Symposium on Cocoa Research (ISCR), Lima, Peru, 13-17 November 2017

The Genomic Selection of *Theobroma cacao*: a new strategy of marker assisted selection to improve breeding efficiency and predict useful traits in new populations

F. Ribeyre (1), O. Sounigo (2), X. Argout (3), C. Cilas (1), B. Efombagn (4), M. Denis (5), J.M. Bouvet (6), O. Fouet (5), C. Lanaud (5)

(1) CIRAD, UPR Bioagresseurs, F-34398 Montpellier, France.
Bioagresseurs, Univ Montpellier, CIRAD, Montpellier, France.
(2) CIRAD, UPR Bioagresseurs, Palmira, Colombia.
Bioagresseurs, Univ Montpellier, CIRAD, Montpellier, France.
(3) CIRAD, UMR AGAP, Palmira, Colombia.
AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.
(4) IRAD, Yaoundé, Cameroon
(5) CIRAD, UMR AGAP, F-34398 Montpellier, France.
AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.
(6) CIRAD, UMR AGAP, 101 Antananarivo, Madagascar.
AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.

Summary

Genomic selection (GS) is a statistical approach that utilizes all available genome-wide markers simultaneously and phenotypic traits of a "training population" to estimate breeding values or total genetic values. For breeding programs, GS is a promising alternative to the traditional marker-assisted selection for manipulating complex polygenic traits often controlled by many small effect genes. A major interest of this method is also to be able to make predictions of trait values, from a training population, on trees only genotyped by molecular markers

The use of the appropriate statistical model remains one of the critical issues of the GS. The relative performance of biometrical models is expected to depend on the genetic background of the traits under assessment.

The objective of this study was to estimate the reliability of different models of genomic selection to predict two agronomic traits of cacao - yield and resistance to *Phytophthora* pod rot.

The study was performed on 287 trees present in a cacao farm plot in Cameroon, belonging to several full-sib progenies released to farmers as commercial varieties.

Each tree was genotyped, using more than 5,000 GBS (genotyping by sequencing) based on SNP markers, and assessed for the mean average of one bean, a trait contributing to cocoa yield, and the % of rotten, as a measure of resistance to Phytophthora megakarya.

Two models were used: Best linear unbiased prediction model and Bayesian lasso model. Cross-validation was used to test their predictive ability. It is an assumption-free method using an estimation set for model training and an independent test set for prediction.

Predictive ability of models was good for both traits indicating that GS is a promising method to improve these cocoa traits. However, it was slightly higher for average weight of a bean (R= 0.59) than for % of rotten pods (R= 0.42).

Introduction

In Cacao (*Theobroma cacao* L.), quantitative traits loci (QTLs) mapping were identified for useful traits such as resistance to pod rot, caused by the oomycete *Phytophthora megakarya*, yield and quality attributes. The identification of multiple QTLs involved in resistance to *Phytophthora* has provided the possibility to improve durability of resistance in cocoa by a possible combination of many different resistance genes located in different chromosome regions using marker assisted selection (Risterucci et al., 2003). The use of a meta QTL analysis approach by Lanaud et al. (2009) has confirmed the existence of several sources of resistance to different diseases of cocoa. However, QTLs were found located in several genomic regions in most of these studies, and the combination of favorable QTL alleles in a same variety becomes complexe when the number of QTLs increases.

In contrast to methods in which markers tagging major effect QTL are detected, genomic selection (GS), proposed by Meuwissen et al. (2001) uses all genetic markers simultaneously to generate a model to predict genomic breeding values (GEBVs), thereby aiming to capture the total additive genetic variance for the trait of interest. It was developed to overcome the limitations of the QTLs, as a solution for the prediction of performance in complex traits. Therefore, GEBVs of some complex traits can be predicted as the sum of all markers effects by regressing phenotypic values on all available markers.

A major interest of this method is to be able to make predictions of trait values, from a training population, on trees only genotyped by molecular markers. The use of the appropriate statistical model remains one of

the critical issues of the GS. The relative performance of biometrical models is expected to depend on the genetic background of traits under assessment. The accuracy of GS, which is the correlation between GEBV and true observed breeding value is affected by several factors: the linkage disequilibrium between markers, the relationships between the training sets and the test sets, the heritability of the trait, the size of the training population, the number of markers, the statistical method to estimate the GEBV, the distribution of underlying QTL effects and the genotype x environment interaction (Crossa et al., 2011; Riedelsheimer et al., 2012; Windhausen et al., 2012).
The objective of this study was to estimate the reliability of different models of genomic selection to predict two agronomic traits of cacao - bean weight and resistance to *Phytophthora* pod rot. More specifically, the purposes were to evaluate and compare the genetic values (GEBV-genomic expected breeding values) predicted by two statistical models and evaluate their accuracy.

Materials and method

Plant Materials
The subset of 287 cocoa trees of the study was selected in a 20-years old commercial cacao plot located in the administrative subdivision of Mbankomo, in the center region of Cameroon, because they yielded more than 17 pods over a three year period.
These trees are issued from seeds released by the extension services, from pods harvested in bi-clonal seed gardens, from natural pollination. A parentage analysis performed on the cocoa trees in the plot allowed Sounigo et al (2016) to identify the two parents of 157 of the trees and to reveal that 21 other trees were issued from selfing.

Phenotyping
Each of the 287 cocoa trees of the study was harvested during 3 years (from April 2007 to March 2010), on a weekly basis and both healthy and rotten pods (figure 1) were counted at each harvest. Studied trait is the total percentage of rotten pods.
A total number of 3755 pods were harvested on 232 cocoa trees for bean assessment. A sample size ranging between 4 and 69 pods per tree was used for measuring the mean weight of one fermented and dried bean. Studied trait is the average weight of a bean.

Figure 1 : Rotten pods (due to *Phytophthora megakarya*) on a cocoa tree

Genotyping
The populations were genotyped with 50224 DArT markers generated by Diversity Arrays Technology Pty Ltd.

Markers selection
Only markers without missing data were selected. Markers with minor allele frequency under 5% were removed.

GEBV estimation
The accuracies of genomic prediction is known to vary depending in particular on the number of genes involved in trait studied. Different models result in different distributions of markers'effect. Two models were used to estimate GEBV using DArT markers: Genomic best linear unbiased prediction (GBLUP) and Bayesian LASSO (BL). GBLUP is a linear mixed model using the kinship matrix between individuals based on genotypes. It assumes a normal distribution of genetic values with a common variance. It supposes that there is a large number of markers with small effects. It doesn't allow calculating individuals markers effects. Bayesian LASSO (BL) assumes a double exponential distribution of markers effects with a marker-specific prior variance for a differential shrinkage of each marker effect (Perez and de los Campos 2014; Perez et al. 2010). Priors were chosen from Crossa et al. (2010). It suppose there is a lot of markers with effects near 0 and some with moderate to large effects.

Accuracy and predictability
Genomic prediction accuracies were measured by the Pearson correlation between GEBVs and the observed phenotypic values. Cross-validation methods were used to evaluate predictive ability of models. Sample data were partitioned into a training set, which is used to build the model and to estimate marker effects, and a validation set, which is used to evaluate the predictive ability of the model. Therefore, half of the genotypes were randomly sampled for the training population. The model is then applied to the 50% others trees used as the validation set and roles were reversed. The process was iterated 30 times to estimate the predictive ability of each fold within each replication calculated as correlation coefficient. For GBLUP model, a fivefold cross validation was assessed, meaning 80% of data were in the training set and 20% in the validation set.

Computations were carried out using R-3.4.1 (R Core Team, 2017) and Synbreed R-package version 0.12-6 (Wimmer et al., 2012).

Results

Percentage of rotten pods varied between 5.4% and 92.5%. *Phytophotora megakarya* causes serious damage since 50% of studied trees have more than 60% of pods attacked (Figure *2*). Average weight of a bean varied between 0.69 and 1.95 grams with a mean of 1.12 (Figure *3*).

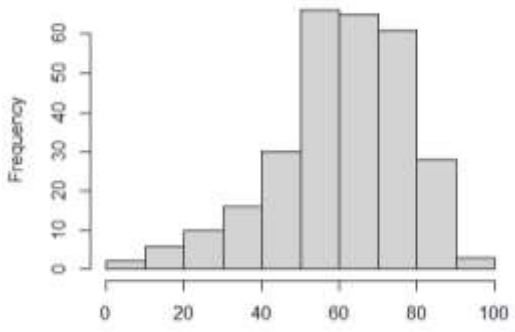

Figure 2: Histogram of the percentage of rotten pods, on 287 trees during 3 years

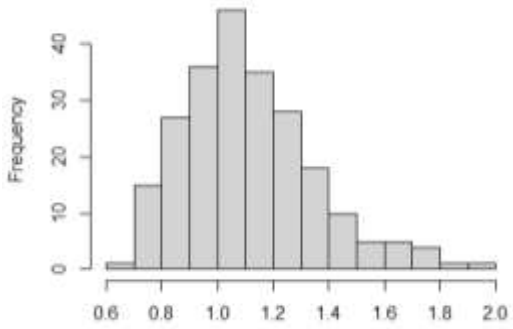

Figure 3: Histogram of the average weight of a bean, on 232 trees

Due to missing values and minor allele frequency under 5%, 12664 and 28636 markers were respectively removed. Thus, models were constructed using 8895 markers without missing data.

Accuracy of predictions were high for both traits and both models. Considering the whole population, correlations between observed percentage of rotten pods and predicted ones reached 0.92 (BL) and 0.90 (GBLUP) (Figure *4*). They were almost similar for average weight of a bean: 0.87 (BL) and 0.90 (GBLUP) (Figure *5*).

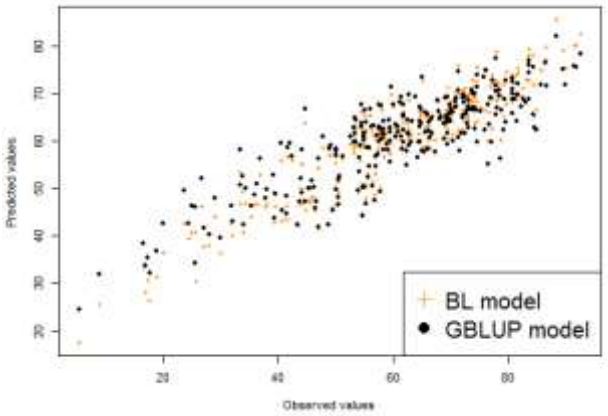

Figure 4: predicted and observed percentages of rotten pods for GBLUP and BL models build on whole data

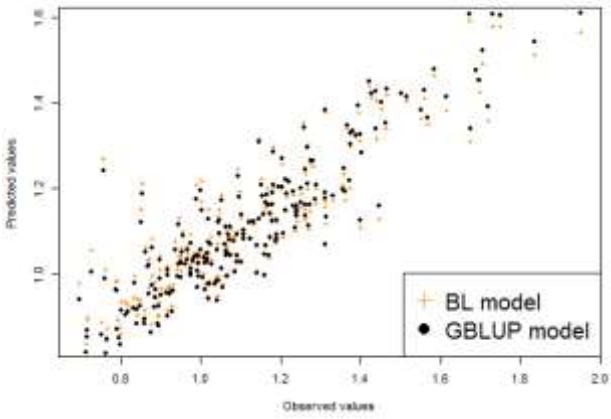

Figure 5: predicted and observed average weight of a bean for GBLUP and BL model build on whole data

Using cross validation, predictive abilities were lower than accuracies (Figure *6* and Figure *7*).

Using 50% of trees to predict percentage of rotten pods in the 50% others, predictive abilities varied among the 30 repetitions from 0.19 to 0.52 with an average of 0.37 for BL model and from 0.28 to 0.52 with an average of 0.42 for GBLUP model. For average weight of a bean, they varied from 0.43 to 0.67 with an average of 0.58 for BL model and from 0.46 to 0.70 with an average of 0.59 for GBLUP model.

Using 80% of trees to predict percentage of rotten pods from 20% others, predictive abilities varied among the 30 repetitions from 0.20 to 0.66 with an average of 0.46 for GBLUP model. For average weight of a bean, they varied from 0.17 to 0.81 with an average of 0.62 for GBLUP model.

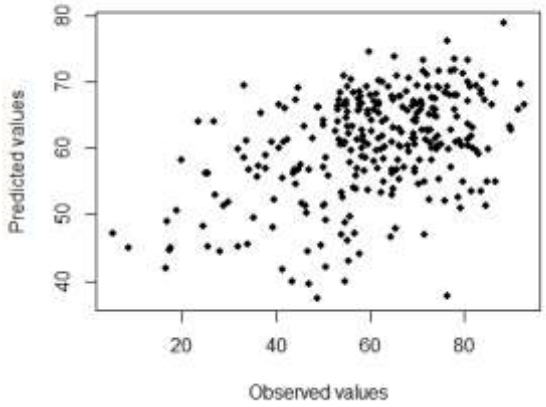

Figure 6: predicted and observed percentage of rotten pods for a repetition of cross-validation using GBLUP model

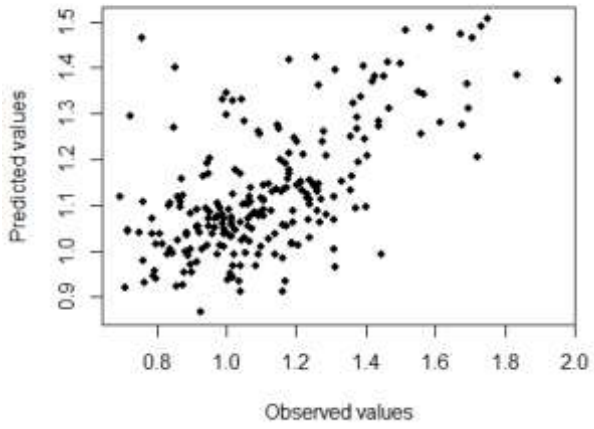

Figure 7: predicted and observed average weight of a bean for a repetition of cross validation using GBLUP model

In order to test the efficiency of genomic selection to select more resistant cocoa trees, we compared the % of rotten pods and the average weight of a bean, observed in the two sets of individuals having the 10% higher or lower predicted values for each trait. (Table *1*).

Table 1: Predicted and observed mean for 10% higher predicted

	Observed mean of 10% higher predicted	Observed mean of 10% lower predicted
Percentage of rotten pods	67	48
Average weight of a bean	1.6	0.96

Discussion

Less than 18% of the markers were considered for the model.

When using the whole data set, accuracy of genomic selection is very high for both traits and both models. Predictive ability, calculated on a part of the population, to predict breeding values on the individuals of the other part of the population was lower. For the average weight of a bean, predictive ability was similar for both models around 0.59. For percentage of rotten pods, it was slightly lower than for bean weight around 0.4. But GBLUP model was higher than BL model (0.42 / 0.37). GBLUP model hypotheses a lot of markers with low effect, whereas BL model hypotheses a lot of markers with 0 effect and some with moderate to large effects. Percentage of rotten pods could be driven by a lot of small markers effects.

Variability of predictive abilities is high between repetitions. The studied trees are issued from several crosses between different progenitors and using more individuals in the training set slightly improve the predictive abilities of both traits (0.42 to 0.46 and 0.59 to 0.62) in the other part of the population. But variability of abilities (interval between minimal and maximal abilities) was higher.

The broad sense heritability of average bean weight is reported as high in literature sometimes exceeding 0.70 (Fallo J. and Cilas C. 1998), while the values for the resistance to Phytophthora in the field were found lower, ranging between 0.42 and 0.51 (Thévenin et al. 2005, Tahi et al. 2006). It is considered as a "Heritable" character unlike the percentage of rotten pods influenced by environment. It could explain why GS model is better to predict the average weight of a bean than the percentage of rotten pods. Nevertheless, for both traits, when considering mean prediction of 10% higher and 10% lower, models give consistent information and allow the selection of individuals improved for resistance to *Phytophthora megakarya* and bean weight.

Predictive ability of models was good for both traits indicating that GS is a promising method to improve these cocoa traits. It was slightly higher for average weight of a bean and it would be interesting to test if taking into account spatial dependence for predicting percentage of rotten pods could improve predictive ability of model. As a perspective, GS could be tested to predict tolerant cocoa trees to disease only present in another environment, as for diseases present only in a particular continent, like whitches'broom and moniliasis.

Références

Crossa, J., de los Campos G., Pérez P., Gianola D., Burgueño J., Araus J.L., Makumbi D., Singh R.P., Dreisigacker S., Yan J., Arief V., Banziger M., and Braun H.-J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724.

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

Fallo J., Cilas C. (1998). Etude génétique de la granulométrie des fèves de cacaoyer (*Theobroma cacao* L.) : relation avec des caractères agronomiques. *Plantations, Recherche, Développement*, 5 (3) : 195-200.

Lanaud C., Fouet O., Clement D., Boccara M., Risterucci A.M., Surujdeo-Maharaj S., Legavre T., Argout X. (2009). A meta-QTL analysis of disease resistance traits of Theobroma cacao L. Molecular Breeding. 24 (4): 361-374.

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

Pérez, P., de los Campos G., Crossa J., and Gianola D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression package in R. Plant Gen. 3:106–116.

Pérez P., de los Campos G. (2014). Genome-wide regression & prediction with the BGLR statistical package Genetics, 198, pp. 482–495

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Riedelsheimer C., Technow F., Melchinger A.E. (2012). Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics 13: 452.

Risterucci A.M., Paulin D., N'Goran J.A.K., Lanaud C. (2003) Identification of QTL related to cocoa resistances to three species of *Phytophtora*. Theor Appl Genet 108:168–174

Sounigo O., Nsouga F., Fouet O., Lanaud C., Efombagn Mousseni I.B. (2016). Combined use of parentage analysis and phenotypic assessment to evaluate the performances of cocoa (Theobroma cacao L.) varieties in farmers' field. International Journal of Plant Breeding and Genetics. Vol. 3 (1), pp. 166-176.

Tahi G.M., Kebe B.I., N'Goran1 J.A.K., Sangare A., Mondeil F., Cilas C., Eskes A.B. (2006). Expected selection efficiency for resistance to cacao pod rot (Phytophthora palmivora) comparing leaf disc inoculations with field observations. Euphytica. 149: 35–44

Thevenin J.M., Umaharan R., Surujdeo-Maharaj S., Latchman B., Cilas C., and Butler D. R. (2005). Relationships between Black Pod and Witches'-Broom disease in Theobroma cacao. Phytopathology. Vol. 95, No. 11: 1301-1307.

Wimmer V., Albrecht T., Auinger H.J. and Schoen C.C. (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics, 28: 2086-2087

Windhausen V.S., Atlin G.N., Hickey J.M., Crossa J., Jannink J.L., Sorrells M.E., Raman B., Cairns J.E., Tarekegne A., Semagn K., Beyene Y., Grudloyma P., Technow F., Riedelsheimer C. and Melchinger A.E. (2012). Effectiveness of Genomic Prediction of Maize Hybrid Performance in Different Breeding Populations and Environments. G3: Genes, Genomes, Genetics November 1, vol. 2 no. 11 1427-1436

Acknowledgments

We acknowledge the Agropolis Fondation for funding.