Guideline for paper in Proceedings and Book of Abstracts

A2

Advances in oil palm genomic selection

David CROS[1,2]

Florence JACOB[3]; Achille NYOUMA[4]; Billy TCHOUNKE[4]; Dadang AFANDI[5]; Indra SYAHPUTRA[5]; Benoit COCHARD[3]

ABSTRACT

More efficient methods are required to breed oil palm (Elaeis guineensis Jacq.) for yield maximization, in order to meet the increased demand for palm oil while limiting environmental impacts. Today, genomic selection (GS) appears to be a disruptive improvement that can speed up breeding schemes by avoiding field trials in some cycles and increase selection intensity by the application of selection to a larger number of candidates than with the current methods. Oil palm is becoming a model species for GS, as it is one of the perennial crops with the largest number of published articles. GS was evaluated in oil palm for the prediction of parental general combining abilities and performances of hybrid crosses and clones. In all cases, GS accuracies high enough to allow selection were obtained for some traits. Best accuracies were obtained when training and validation populations were highly related, such as full-sibs or progenies. Array-based SNPs and GBS-derived SNPs allowed cost effective GS predictions, with densities of a few thousand markers being sufficient. Widely used statistical methods of GS predictions GBLUP and rrBLUP appeared efficient, and could be optimized by SNP filtering methods. Approaches to limit the increase in the rate of inbreeding associated with GS were identified. Evaluations of the annual genetic progress showed that GS should bring it to an unprecedented level. Further studies remain required for the optimal application of GS in oil palm. They should focus in particular on the optimization of training populations, the improvement of prediction models, the variation of GS accuracy between families, the use of multi-omics data (transcriptomics, proteomics, etc.), the modeling of G × E interactions and inter-specific selection.

[1] CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP, F-34398 Montpellier, France.
[2] AGAP, Université de Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France
[3] PalmElit SAS, 34980 Montferrier sur Lez, France
[4] University of Yaounde 1, Yaounde, Cameroon
[5] P.T. SOCFINDO Medan, Medan 20001, Indonesia

INTRODUCTION

For quantitative traits (i.e. complex traits under the control of a large number of genes with small effects), the efficiency of QTL-based marker assisted selection (MAS) is limited (Muranty et al. 2014; Grattapaglia et al. 2018), because it overestimates the effect of strong QTLs and fails to exploit weak QTLs. Genomic selection (GS, Meuwissen et al. 2001) was developped to address these problems. It is now largely used in animal breeding, particularly in dairy cattle where it has doubled the rate of the genetic gain (Wiggans et al. 2017). It is also progressively incorporated in plant breeding, that it should make significantly more efficient (Varshney et al. 2017). GS predicts the genetic value of selection candidates, usually with unknown genetic performance. It uses specific statistical methods, such as BLUP, and analyzes jointly a large number of markers spread along the whole genome. It uses the genotypic and phenotypic data of a training (or calibration) population and a mixed model that can predict the additive genetic value (GEBV, genomic estimated breeding values) or the total genetic value (i.e. including the non-additive effects) of the selection candidates. GS can therefore reduce phenotyping, thus shortening the breeding cycle and/or allowing applying selection to a larger number of candidates (i.e. increasing selection intensity).

GS efficiency is assessed by its selection accuracy (*r*), i.e. the correlation between the genetic value estimated with the genomic model (GEGV) and the true genetic value (TGV) of individuals constituting the validation population. In practice TGVs are unknown and GS is evaluated on its prediction accuracy, which is the correlation between GEGVs and an estimate of TGVs, obtained with the phenotypic data available on the validation individuals (usually their own phenotypes or the phenotypes of their progenies). The difference between selection accuracy and prediction accuracy depends on the reliability of the estimate of the TGVs. Many factors affect GS accuracy, including marker type and density, distribution of QTL effects, linkage disequilibrium between markers and QTLs, training population size, relationship between training and selection populations, trait heritability and statistical methods of prediction. As in other species, GS accuracy in oil palm is usually estimated by cross-validation at a single experimental site (Cros et al. 2015b; Kwong et al. 2017a, b) or by between-site validation (Cros et al. 2017). However, single-site cross-validations may overestimate accuracy, and it is preferable to have at least two sites to evaluate GS (Lorenz et al. 2011).

The potential of GS for oil palm breeding has already been investigated in several studies conducted by various research groups (see Nyouma et al. 2019a for a review), making oil palm one of the perennial crops with the largest number of published articles about GS. Here we will summarized the results obtained so far, presenting, first, how various factors affect the GS accuracy and, second, what can be expected in terms of rate of genetic progress.

INFORMATION CAPTURED BY MARKERS

Using SNPs and without optimizing the training and validation populations, prediction accuracies ranging from 0.14 and 0.73 were obtained for various yield components, showing the ability of genomic models to predict the genetic value of unevaluated selection candidates (Cros et al. 2017; Kwong et al. 2017a, b). In particular, for five yield components (FFB, O/M, BN, BW and M/F), the GS model predicted the performance of unevaluated hybrid crosses more accurately than a control model using pedigree data instead of markers (Cros et al. 2017). This showed the ability of GS to capture genetic differences within full-sib families (i.e. Mendelian segregation terms) in addition to genetic differences between families, enabling the selection of the best individuals of the best families, as currently done with the phenotypic breeding schemes. The same conclusion was reached in Kwong et al. (2017b), where GS prediction accuracies above zero, ranging from 0.18 to 0.47, were obtained in a GS evaluation

considering a single full-sib family. Similarly, Cros et al (2015b) obtained GS prediction accuracies above 0.5 within full-sib families. However, the latter study also showed that GS could also, depending on trait and population, fail to capture Mendelian segregation. In this case, GS predictions only revealed, at best, between-family differences.

MOLECULAR DATA

The first empirical studies in oil palm were made with SSRs (simple sequence repeats) (Cros et al. 2015b; Marchal et al. 2016). However, oil palm GS studies now use single nucleotide polymorphisms (SNP), from SNP arrays (Kwong et al. 2016, 2017a, b; Ithnin et al. 2017) or genotyping by sequencing (GBS) (Cros et al. 2017; Nyouma et al. 2019b). SNPs are needed as the practical application of GS requires a high throughput genotyping approach, as the number of individuals to genotype is large. Also, the use of SNPs allows reaching higher densities, leading to higher accuracies. Thus, Kwong et al. (2017b) using 200K SNPs obtained mean GS prediction accuracies of 0.31 over palm oil yield components, against 0.21 with 135 SSRs.

Several studies in oil palm showed that, although the marker density required to reach the maximum GS accuracy was affected by marker type, marker sampling, trait and population, a few thousands SNPs were enough (Marchal et al. 2016; Kwong et al. 2017a; Cros et al. 2017; Nyouma et al. 2019b). This marker density is low compared to the densities generally used in other species, which results from the high rate of inbreeding in oil palm breeding populations. To increase the cost efficiency of GS, SNP filtering can be used to reduce the marker density, leading to accuracies equal or higher than the accuracies obtained with all the SNPs. This can be done by using the SNPs with the highest association scores estimated in a genome-wide association study (Kwong et al. 2017a) or, with GBS, using the SNPs with less than 5% missing data (Cros et al. 2017).

The GS statistical models are not able to handle missing molecular data, which therefore must be imputed, i.e. replaced by the most likely genotypes. The percentage of missing data is very low with SNP arrays (< 1% in Kwong et al. (2016)) and SSRs (< 3% in Cros et al. (2015b)), but they reach significant levels with GBS (13.2% in Cros et al. (2017)). In this case, the imputation approach is likely of importance. Many imputation methods are available (Wang et al. 2016), and comparing their efficiency in oil palm would be of interest. So far, the only study considering this aspect showed that, when using the Beagle software, taking pedigree information into account for imputation improved GS accuracy (Cros et al. 2017).

TRAINING AND APPLICATION POPULATIONS

As in other species, GS accuracy in oil palm is strongly affected by the relationship between training and application individuals (Cros et al. 2015b). Implementing GS in full-sibs or progenies of the training individuals would therefore maximize the efficiency of the approach.

The training and application populations can also be optimized. Several approaches have been developed for this purpose. However, only the CDmean criterion was tested so far on oil palm data, in a study that showed that it allowed better defining the training population, through the optimized sampling of individuals to phenotype among a set of genotyped individuals (Cros et al. 2015b).

Another way to increase GS accuracy is to increase the size of the training set. This can be done for example by aggregating data from consecutive breeding cycles. Simulations in oil palm showed that the use of data from two cycles increased the per cycle response to selection by more than 10%, mostly as a result of higher selection accuracy (Cros et al. 2018), and despite the associated reduction in relationship between training and application populations.

STATISTICAL METHODS OF PREDICTIONS

Some genomic predictions methods estimate an effect associated with each marker, while others give the genetic values directly without estimating marker effects (Wang et al. 2018). First, among the prediction methods that estimate an effect for each marker, some methods consider that marker effects are sampled according to a normal distribution common to all markers, which is relevant for traits following the infinitesimal model. This is the case of random regression BLUP (RR-BLUP) and Bayesian random regression (BRR). However, as the genetic determinism of some quantitative traits may also include loci with strong effects, other methods allowing marker specific genetic variances were developed: Bayes A, Bayes B, Bayes Cπ, Bayes Dπ, Bayesian LASSO, etc. Second, the most common method to estimate GEGV directly is the genomic best linear unbiased predictor (GBLUP). It uses the genomic information to compute the relationship matrix, which accounts for the random sampling of alleles at meiosis (Mendelian sampling) and thus gives realized relationships, making it possible to obtain the GEGV of unevaluated individuals.

A wide range of statistical methods has been applied for genomic predictions in oil palm, and comparisons showed that they did not significantly affect GS accuracy (Cros et al. 2015b; Kwong et al. 2017b; Ithnin et al. 2017). This suggests that the traits considered in oil palm GS studies so far (i.e. yield components, bunch analysis traits and vegetative parameters) are highly polygenic and follow the infinitesimal model. Also, this shows that the GBLUP and RR-BLUP methods, that are widely used for GS predictions due to their simplicity and computational efficiency, are suitable for oil palm.

DATA MODELING

Various modeling approaches have been used for genomic predictions in oil palm. Some studies applied GS models independently in each parental group. In this case, some authors used data records consisting in parental performances in crosses with the other group, i.e. GCAs (Cros et al. 2015b) or testcross phenotypic means (Wong and Bernardo 2008), and parent genotypes. By contrast, Ithnin et al. (2017) and Kwong et al. (2017b) used parental phenotypes as data records. In theory, as the goal is to develop hybrid cultivars, this might be not optimal for some traits, as parental phenotypes may not reflect performance in hybrid crosses due to gene-frequency differences between parental populations and non-additive effects. Other studies applied GS models that jointly predicted the GEBV of the parents of the two heterotic groups (Cros et al. 2015a, 2017, 2018; Marchal et al. 2016). So far, a comparison of these different modeling approaches is lacking.

Kwong et al. (2016) obtained a GS prediction accuracy of 0.65 using a population comprising hybrid and parent individuals (Deli and group B). Although such an accuracy is high enough for breeding purposes, this type of complex population could possibly give greater GS accuracy if analyzed with a model designed to jointly consider parental and hybrid data, like the one applied in pigs by Vitezica et al. (2016).

Another approach investigated by simulation in oil palm consisted in training the GS model using molecular data of individual hybrids taking into account the parental origin of marker alleles (Cros et al. 2015a). This gave higher GS accuracies than using only parental genotypes. Kwong et al. (2017a) used molecular data of individual hybrids in an empirical study, but did not consider the parental origin of alleles. Nyouma et al. (2019b), when predicting the genetic value of candidate ortets, showed that the best approach differed among traits and SNP density, i.e. that taking into account the parental origin of marker alleles led to the highest GS accuracy in some cases, while for other cases it was more efficient to ignore

this information. The usefulness of modeling the parental origin of marker alleles in oil palm hybrids should be further investigated.

ANNUAL GENETIC PROGRESS

GS accuracy is a key parameter to evaluate GS as it is directly related to the annual genetic progress (R), according to $R=r\times i\times\sigma_g/L$, where i is the selection intensity, σ_g the genetic variance and L the number of years per breeding cycle. However, the comparison of GS and conventional selection must take into account their respective values for r, i and L. Indeed, even if r is lower in GS than in the conventional phenotypic evaluations, GS can still increase R if it allows a sufficient decrease in L and/or increase in i.

Wong and Bernardo (2008) started a simulation study with an initial breeding population derived from the selfing of a hybrid, followed by two cycles of conventional breeding. At each cycle, the breeding population was crossed with a tester to allow phenotypic selection for yield performance, and the selected individuals were crossed to produce the following generation. With QTL-based MAS and GS, the initial population was also genotyped and used to estimate marker effects, and in the following cycles, selection was made on markers. This reduced the length of the breeding cycles and enabled three consecutive selection cycles, with a total number of years over the four cycles equivalent to the two cycles in conventional phenotypic selection. GS and conventional selection outperformed QTL-based MAS in terms of selection response, while GS outperformed conventional selection when the population size reached 50 to 70 individuals, and then increased selection response by 4% to 25%, depending on population size, heritability and number of QTLs.

In another simulation study, conventional RRS and GS were compared over four cycles (Cros et al. 2015a). With GS, the cycles including hybrid progeny tests were used to train a model applied to make a selection among unevaluated individuals of the same cycle (i.e. sibs of the evaluated individuals) and/or of the following generations. The simulation quantified the effect of three parameters on the annual selection response: frequency of progeny tests (from model training only in first cycle to training in every cycle), the number of GS candidates (120 and 300) and GS strategy (genotyping limited to the parents of the calibration hybrids [RRGS_PAR] or also genotyping hybrid individuals [RRGS_HYB]). It showed that GS could increase annual genetic progress by reducing the generation interval and by increasing the selection intensity, despite the fact that GS accuracy for unevaluated hybrid parents was lower than the accuracy of progeny tested parents. Among the strategies evaluated, RRGS_HYB with the genotyping of 1,700 hybrid individuals, model training only in the first generation and 300 selection candidates per population and generation was the most efficient, leading to 72% higher annual genetic progress than RRS. Additionally, RRGS_PAR with model training every two generations and 300 selection candidates was shown to be an interesting alternative as, although its genetic progress was lower (46% higher than RRS), it had a lower variability of genetic progress, reduced cost and slower increase in inbreeding over cycles in the parental populations. The authors later studied the effect of aggregating the data of two consecutive cycles to train the RRGS_PAR model and showed that this increased the selection accuracy, leading to an annual genetic progress 37.6% to 57.5% higher than RRS, depending on the number of GS candidates (Cros et al. 2018).

These simulations promise a revolution in the genetic improvement of oil palm. However, even if the empirical studies showed that GS accuracies could be high, they also revealed that so far GS was not efficient for all the traits. Thus, for some traits the GS model gave low prediction accuracies (<0.2) and/or did not predict the genetic value of unevaluated individuals better than a control model using pedigree data instead of markers (Cros et al. 2015b, 2017, Ithnin et al. 2017, Nyouma et al. 2019). Yet, the simulations showed that the main

advantage of GS was its ability to shorten the breeding cycles by avoiding field evaluations in some cycles, which is only possible if GS is efficient for all the traits that are currently subjected to phenotypic selection. Otherwise, the field trials remain necessary in all cycles. Therefore, the practical application currently envisaged to start implementing GS in oil palm is a two-stage scheme, with an initial stage of genomic selection prior to field trials. For the selection of parents of hybrid crosses, this would be better than the current first stage of phenotypic selection for two reasons. First, the number of yield components for which GS is efficient is greater than the number of traits currently subjected to phenotypic preselection. Second, the current selection prior to progeny tests is made on the parental phenotypes, even though they may be poor indicators of performance in hybrid crosses, while genomic predictions could be obtained with a model calibrated on hybrid phenotypes. The potential of genomic preselection was quantified based on the GS accuracies empirically obtained by between-site validation for bunch production (Cros et al. 2017), and it showed that this could increase the performance of the selected hybrids by more than 10% compared to a method without preselection, thanks to higher selection intensity. Considering the selection of ortets, Nyouma et al. (2019b) showed that combining genomic predictions and conventional phenotypic evaluations would allow preselecting ortet candidates before clonal trials on all yield components, against currently on only one or two traits with high heritability. This would increase selection intensity and therefore genetic progress.

Simulations showed that GS in oil palm would result in a faster increase in inbreeding in the parental populations than conventional breeding. This could affect negatively seed production, with for example germination problems that could arise from inbreeding depression. This could also negatively impact the genetic progress over the long term. However, approaches of inbreeding management exist and are efficient in the context of oil palm GS, in particular mate selection (Tchounke et al., *in prep*).

To be applied in practice, GS must also be cost efficient. Although GS generates additional costs due to genotyping, these costs are low in comparison to the cost of phenotyping. Thus, Jacob et al. (2017) indicated that, even assuming a genotyping cost per sample as high as 300€, which seems to be the maximum possible price for a 300K SNP array, the ratio of genotyping/phenotyping costs lays below 1/20. In addition, these extra costs could possibly be offset in the future by a reduction in phenotyping costs in the GS scheme. In this case, Wong and Bernardo (2008) found that with a genotyping cost of US$0.15 per datapoint, corresponding to genotyping prices for SNPs, the cost per genetic progress unit was 35% to 65% lower with GS than with conventional selection.

CONCLUSION

GS has the potential to speed up oil palm genetic progress to a previously unprecedented level by avoiding field trials in some cycles and by predicting the genetic value of a much larger number of selection candidates. Further studies remain necessary to optimize oil palm GS. They should focus in particular on the design of the training population, the statistical model used for predictions, new traits (for instance resistance to diseases), the use of multi-omics data, G × E interactions, variation of GS accuracy between families and inter-specific selection.

REFERENCES

Cros D, Bocs S, Riou V, et al (2017) Genomic preselection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses. BMC Genomics 18:839. doi: 10.1186/s12864-017-4179-3

Cros D, Denis M, Bouvet J-M, Sanchez L (2015a) Long-term genomic selection for heterosis without dominance in multiplicative traits: case study of bunch production in oil palm. BMC Genomics 16:651.

Cros D, Denis M, Sánchez L, et al (2015b) Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). Theor Appl Genet 128:397–410. doi: 10.1007/s00122-014-2439-z

Cros D, Tchounke B, Nkague-Nkamba L (2018) Training genomic selection models across several breeding cycles increases genetic gain in oil palm in silico study. Mol Breed 38:89. doi: 10.1007/s11032-018-0850-x

Grattapaglia D, Silva-Junior OB, Resende RT, et al (2018) Quantitative Genetics and Genomics Converge to Accelerate Forest Tree Breeding. Front Plant Sci 9:1693. doi: 10.3389/fpls.2018.01693

Ithnin M, Xu Y, Marjuni M, et al (2017) Multiple locus genome-wide association studies for important economic traits of oil palm. Tree Genet Genomes 13:103. doi: 10.1007/s11295-017-1185-1

Jacob F, Cros D, Cochard B, Durand-Gasselin T (2017) Agrigenomics in the breeder's toolbox: latest advances towards an optimal implementation of genomic selection in oil palm. In: International Seminar on 100 Years of Technological Advancement in Oil Palm Breeding & Seed Production. ISOPB conference, 13 November 2017, KLCC, Kuala Lumpur, p. 21.

Kwong QB, Ong AL, Teh CK, et al (2017a) Genomic Selection in Commercial Perennial Crops: Applicability and Improvement in Oil Palm (Elaeis guineensis Jacq.). Sci Rep 7:2872. doi: 10.1038/s41598-017-02602-6

Kwong QB, Teh CK, Ong AL, et al (2016) Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm. Mol Plant 9:1132–1141. doi: 10.1016/j.molp.2016.04.010

Kwong QB, Teh CK, Ong AL, et al (2017b) Evaluation of methods and marker Systems in Genomic Selection of oil palm (Elaeis guineensis Jacq.). BMC Genet 18:107. doi: 10.1186/s12863-017-0576-5

Lorenz AJ, Chao S, Asoro FG, et al (2011) Genomic Selection in Plant Breeding: Knowledge and Prospects. In: Donald L. Sparks (ed) Advances in Agronomy. Academic Press, , p. 77–123.

Marchal A, Legarra A, Tisné S, et al (2016) Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests. Mol Breed 36:1–13. doi: 10.1007/s11032-015-0423-1

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

Muranty H, Jorge V, Bastien C, et al (2014) Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. Tree Genet Genomes 1–20. doi: 10.1007/s11295-014-0790-5

Nyouma A, Bell JM, Jacob F, Cros D (2019a) From mass selection to genomic selection: one century of breeding for quantitative yield components of oil palm (Elaeis guineensis Jacq.). Tree Genet Genomes 15:69. doi: 10.1007/s11295-019-1373-2

Nyouma A, Jacob F, Riou V, et al (2019b) Prédiction de la valeur génétique clonale chez le palmier à huile à partir de données génomiques haute densité et du modèle linéaire mixte. Dakar, Sénégal / 28-30 avril 2019

Varshney RK, Roorkiwal M, Sorrells ME (2017) Genomic Selection for Crop Improvement, 1st edn. Springer International Publishing, Cham, Switzerland

Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: Current status and prospects. Crop J 6:330–340. doi: 10.1016/j.cj.2018.03.001

Wang Y, Lin G, Li C, Stothard P (2016) Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle. Springer Sci Rev 4:79–98. doi: 10.1007/s40362-017-0041-x

Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS (2017) Genomic Selection in Dairy Cattle: The USDA Experience. Annu Rev Anim Biosci 5:309–327. doi: 10.1146/annurev-animal-021815-111422

Wong CK, Bernardo R (2008) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theor Appl Genet 116:815–824. doi: 10.1007/s00122-008-0715-5