# Multivariate genomic model improves analysis of oil palm (Elaeis guineensis Jacq.) progeny tests

Alexandre Marchal, Andres Legarra, Sebastien S. Tisne, Catherine Carasco-Lacombe, Aurore Manez, Edyana Suryana, Alphonse Omore, Bruno Nouy, Tristan Durand-Gasselin, Léopoldo Sanchez, et al.

Molecular Breeding, 2016, 36(1), 36:2 http://dx.doi.org/10.1007/s11032-015-0423-1

# Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests

Alexandre Marchal, Andrés Legarra, Sébastien Tisné, Catherine Carasco-Lacombe, Aurore Manez, Edyana Suryana, Alphonse Omoré, Bruno Nouy, Tristan Durand-Gasselin, Leopoldo Sánchez, Jean-Marc Bouvet, David Cros

*A. Marchal, S. Tisné, C. Carasco-Lacombe, A. Manez, J.M. Bouvet, D. Cros* (✉)
*CIRAD, UMR AGAP (Genetic Improvement and Adaptation of Mediterranean and Tropical Plants Research Unit), 34398 Montpellier, France*
*e-mail: david.cros@cirad.fr*

*Andrés Legarra*
*INRA, UMR1388, GenPhySe, 31326 Castanet Tolosan, France*

*E. Suryana*
*P.T. SOCFINDO Medan, Medan 20001, Indonesia*

*A. Omoré*
*INRAB, CRAPP, Pobè, Benin*

*B. Nouy, T. Durand-Gasselin*
*PalmElit SAS, 34980 Montferrier sur Lez, France*

*L. Sánchez*
*INRA, UR0588, UAGPF (Forest Tree Improvement, Genetics and Physiology Research Unit), 45075 Orléans, France*

Molecular Breeding, 2016, 36(1), 36:2 http://dx.doi.org/10.1007/s11032-015-0423-1

**Abstract**

Genomic selection is promising for plant breeding, particularly for perennial crops. Multivariate analysis, which considers several traits jointly, takes advantage of the genetic correlations to increase accuracy. The aim of this study was to empirically evaluate the potential of a univariate and multivariate genomic mixed model (G-BLUP) compared to the traditional univariate pedigree-based BLUP (T-BLUP) when analyzing progeny tests of oil palm, the world major oil crop.

The dataset comprised 478 crosses between two heterotic groups A and B with 140 and 131 parents, respectively, genotyped with 313 SSR. The traits were bunch number and average bunch weight.

We found that G-BLUP with a genomic matrix based on a similarity index had a higher likelihood than T-BLUP. Also, multivariate G-BLUP improved the accuracy of additive effects (breeding values or general combining abilities, GCAs), in particular for the less heritable trait, and of dominance effects (specific combining abilities, SCAs). The average increase in accuracy was 22.5% for GCAs and 18.7% for SCAs. Using 160 markers in group A and 90 in group B was enough to reach maximum GCA prediction accuracy. The contrasted history of the parental groups likely explained the higher benefit of G-BLUP over T-BLUP for group A than for group B.

Finally, G-BLUP should be used instead of T-BLUP to analyze oil palm progeny tests, with a multivariate approach for correlated traits. G-BLUP will allow breeders to consider SCAs in addition to GCAs when selecting among the progeny-tested parents.

**Keywords** *Elaeis guineensis*, genomic selection, multivariate model, empirical data, reciprocal recurrent selection, accuracy

## 1. Introduction

Oil palm (*Elaeis guineensis* Jacq.) is the main oil crop in the world. It bears fruit bunches all year long, and palm oil is extracted from the mesocarp of the fruits. Bunch production is a key component of oil yield, and results from the product of two negatively correlated traits, bunch number (BN) and average bunch weight (ABW) (Corley and Tinker 2003). Commercial oil palms are hybrids between two heterotic groups called A and B. Group A is mostly made up of the Deli population (Asia) and group B of various African populations. Group A palms have a few heavy bunches whereas group B palms have many small bunches, resulting in heterosis of bunch yield in A × B hybrids. This led to the choice of a reciprocal recurrent selection (RRS) breeding scheme in the 1950s (Gascon and de Berchoux 1964; Meunier and Gascon 1972). RRS involves progeny tests in which group A and group B parents are crossed to estimate their general combining ability (GCA), i.e. half their breeding value in hybrid crosses, for each yield component, from the phenotype of their hybrid progenies. So far, parental GCAs are obtained using an univariate mixed-model analysis (i.e. considering one trait at a time) taking pedigree information into account (Soh 1994; Purba et al. 2001). The accuracy of the GCAs (i.e. the correlation between the estimated and the true GCAs) is high, reaching around 0.9 for all yield components (Cros et al. 2015b). However, the progeny tests require a long generation interval (around 20 years) and low selection intensity (less than 200 individuals tested per parental group and generation).

Genomic selection (GS) aims to predict genetic values of candidate individuals. In particular, GS can be applied on candidate individuals without data records, by using their genotype with high density molecular markers and a model calibrated with a training set made of individuals with records and marker data (Meuwissen et al. 2001). GS is then particularly promising when traditional breeding requires extensive phenotyping, like progeny tests, as in this case GS makes it possible to reduce the generation interval and to increase selection intensity. The potential of GS is particularly high for perennial crops (Grattapaglia 2014; Isik 2014; van Nocker and Gardiner 2014). In oil palm, previous studies showed that GS could allow selecting individuals without progeny tests (Wong and Bernardo 2008; Cros et al. 2015b). However, GS also has the potential to improve the analysis of progeny tests. So far, no empirical study has investigated whether GS can increase the accuracy of the GCA of progeny-tested oil palms.

The GS model G-BLUP (VanRaden 2007; Habier et al. 2007) is a mixed model that makes use of molecular information through a genomic matrix (**G**) of realized relationships. Multivariate analysis using mixed modeling (i.e. considering several traits jointly) aims to take advantage of the genetic correlation between traits to increase accuracy (Gilmour et al. 2009). Simulations have shown that multivariate G-BLUP can yield higher

accuracy than univariate G-BLUP, depending on the heritability of the traits ($h^2$) and their genetic correlation (Calus and Veerkamp 2011; Jia and Jannink 2012; Guo et al. 2014). When considering two traits with different $h^2$, bivariate G-BLUP led to a higher increase in accuracy for the trait with the lowest $h^2$. In addition, Jia and Jannink (2012) and Calus and Veerkamp (2011) found that the stronger the genetic correlation, the greater the benefit of using a bivariate G-BLUP. However, Jia and Jannink (2012) did not provide evidence for improved accuracy with multivariate analysis when they used empirical data. In oil palm simulations, Cros et al (2015a) used multivariate models but did not make comparisons with univariate models.

The aim of the present study was to compare the potential of univariate and bivariate G-BLUP with that of the current univariate pedigree based BLUP for the analysis of BN and ABW traits in oil palm progeny tests using real data.

## 2. Material and methods

### Experimental population and phenotypes

The progeny test involved 146 group A parents crossed with 155 group B parents according to an incomplete factorial design with 478 crosses. The crosses were planted between 1995 and 2000 in 26 trials located in the same area, at the Aek Loba estate (SOCFINDO, North Sumatra). All the vegetal material belonged to the PalmElit breeding program (www.palmelit.com). Annual bunch production data, i.e. bunch number (BN) and average bunch weight (ABW), were collected on 30,872 progeny palms of type tenera (thin-shelled commercial type) from 6 years old up to 11 years old. More details on the experimental design are given in Cros et al. (2015b). Phenotypic correlation between ABW and BN was -0.682. The narrow-sense heritabilities $h^2$ of ABW and BN varied with the parental population, $h^2_{BN}$ was higher than $h^2_{ABW}$ in A ($h^2_{BN}$ = 0.31 ± 0.04 [s.e.], $h^2_{ABW}$ = 0.23 ± 0.04) and lower than $h^2_{ABW}$ in B ($h^2_{BN}$ = 0.5 ± 0.05, $h^2_{ABW}$ = 0.57 ± 0.04) (Cros et al. 2015b).

### Molecular data

Among the progeny-tested parents, 140 group A and 131 group B individuals were genotyped. Supplementary Table S1 lists the distribution of these individuals among the populations constituting the parental groups. Genotyping was performed with 313 simple sequence repeat markers (SSR) (Billotte et al. 2005; Tranbarger et al. 2012; Zaki et al. 2012). Phenotypic observation of the fruit type (i.e. shell thickness) was included as a two-allele marker, corresponding to the Sh gene (Singh et al 2013). Missing data (1.7% in group A and 2.9% in group B) were imputed with BEAGLE 3.3.2 (Browning and Browning 2007). Finally, group A had 265 polymorphic SSR (mean 3.05 alleles ± 0.89 (standard deviation)), and group B had 289 polymorphic SSR (mean 6.25 alleles ± 2.35). For each group only the polymorphic markers were used for the genomic models.

### Prediction models

*Univariate T-BLUP*

The traditional pedigree-based mixed model or T-BLUP was used to predict the genetic effects, i.e. the general combining abilities (GCAs) in A x B crosses of the progeny-tested parents and the specific combining abilities (SCAs) of the crosses (dominance effects). The model was:

$$\mathbf{P} = \mathbf{X\beta} + \mathbf{Wb} + \mathbf{Z_A g_A} + \mathbf{Z_B g_B} + \mathbf{Z_D s} + \mathbf{e} \qquad [1]$$

where **P** is the vector of hybrid phenotypes (BN or ABW), **X** and **W** are incidence matrices of the experimental design effects, **β** and **b** are the vectors of fixed and random effects due to the experimental design, respectively,

$\mathbf{Z_A}$, $\mathbf{Z_B}$ and $\mathbf{Z_D}$ are incidence matrices of the genetic random effects, $\mathbf{g_A}$ and $\mathbf{g_B}$ are the vectors of GCA of parents A and B, respectively, **s** is the vector of SCA of crosses, and **e** is the vector of residual effects.

The random genetic effects followed the model of Stuber and Cockerham (1966), with $\mathbf{g_A} \sim N(0, \sigma^2_{g_A} \times \mathbf{A_A})$, $\mathbf{g_B} \sim N(0, \sigma^2_{g_B} \times \mathbf{A_B})$ and $\mathbf{s} \sim N(0, \sigma^2_s \times \mathbf{D})$, where $\sigma^2_{gA}$ and $\sigma^2_{gB}$ are the additive variances of the A and B parents in A × B hybrid crosses, respectively, and $\sigma^2_s$ is the variance of the dominance effects in the A × B population. Given the hybrid nature of the crosses, the **A** matrices contain Malécot's coefficient of coancestry *f* (Malécot 1948), such as $\mathbf{A_{xy}} = \{ f_{xy} \}$ between individuals *x* and *y*. They were built from the pedigrees with the R package synbreed (Wimmer et al. 2012). The **D** matrix is the dominance coancestry matrix between crosses, obtained as $\mathbf{D} = \mathbf{A_A} \otimes \mathbf{A_B}$ [2], i.e. with elements $\mathbf{D}_{AB,A'B'} = f_{AA'}f_{BB'}$, as A and B individuals are unrelated (Stuber and Cockerham 1966; Lynch and Walsh 1998).

Fixed effects were: overall mean, "trial" (26 levels), "block" (152 levels) and "age" (6 levels). Random effects associated with the experimental design were "elementary plots" (3,464 levels), "individual" (30,872 levels), interaction "age*cross" ("**α*s**", 2,855 levels); with "individual" nested in "elementary plots", "elementary plots" nested in "block", and "block" nested in "trial". The random experimental design effects followed a normal distribution of the form $N(0, \sigma^2 \times \mathbf{I})$, where **I** is the identity matrix and $\sigma^2$ the associated variance, with the exception of **α*s** that followed $N(0, \sigma^2_{\alpha*s} \times \mathbf{I}_{6\times6} \otimes \mathbf{D})$. The errors **e** followed $N(0, \sigma^2_e \times \mathbf{I}_{180872\times180872})$, where $\sigma^2_e$ is the residual variance.

Variance parameters were estimated by restricted maximum likelihood (REML) and solutions of the mixed model were obtained by resolving Henderson's mixed model equations (Henderson 1975), using R-ASReml (Gilmour et al. 2009; R Core Team 2014).

*Univariate G-BLUP*

In the genomic selection model G-BLUP, the pedigree coancestry matrices used in [1] were replaced by additive genomic coancestry matrices $\mathbf{G_A}$ and $\mathbf{G_B}$ for groups A and B, respectively.

As some progeny-tested individuals were not genotyped, their pedigree coancestry had to be combined with the molecular coancestry of the genotyped individuals. For this purpose, we used the single-step approach with matrices $\mathbf{H_A}$, $\mathbf{H_B}$ and $\mathbf{D_H}$ designed to combine both genomic and pedigree information (Legarra et al. 2009; Christensen and Lund 2010). For each parental group, **H** inverse was built as follows:

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix}$$

where **A** is the pedigree coancestry matrix including all the individuals in the group, $\mathbf{A_{22}}$ and **G** are the pedigree and the genomic coancestry matrices, respectively, containing only the genotyped individuals. Then $\mathbf{H_D}$ was built in the same way as in equation [2]:

$$\mathbf{D_H} = \mathbf{H_A} \otimes \mathbf{H_B}$$

This led to $\mathbf{g_A} \sim N(0, \sigma^2_{gA} \times \mathbf{H_A})$, $\mathbf{g_B} \sim N(0, \sigma^2_{gB} \times \mathbf{H_B})$ and $\mathbf{s} \sim N(0, \sigma^2_s \times \mathbf{D_H})$.

Three different genomic additive coancestry matrices **G** were compared: $\mathbf{G_{AIS}}$, $\mathbf{G_{OF}}$ and $\mathbf{G_N}$. $\mathbf{G_{AIS}}$ used a similarity index (Lynch 1988; Li et al. 1993) and was defined as:

$$\mathbf{G_{AIS}} = \frac{\mathbf{ZZ^t}}{4L}$$

where **Z** is the genotypic matrix with as many columns as alleles, with the individuals in rows, and containing in the i$^{th}$ column the number of copies of the i$^{th}$ allele ($\mathbf{Z}_{xy} \in \{0,1,2\}$), and L is the total number of markers. This index estimates coancestry from alike-in-state (AIS) alleles, and assumes that each allele was unique in the founder population that generated the population under study (Eding and Meuwissen 2001).

$\mathbf{G_{OF}}$ was obtained according to VanRaden (2007; 2008), with a modification for multiallelic markers:

$$\mathbf{G_{OF}} = \frac{(\mathbf{Z}-\mathbf{P})(\mathbf{Z}-\mathbf{P})^t}{4\sum_{l=1}^{L}(1-\sum_l \sum_a p_{l_a}^2)}$$

where **P** is a matrix containing twice the observed allelic frequency (OF) of the i$^{th}$ allele in the genotyped individuals in the i$^{th}$ column.

The coancestry matrix of VanRaden (2007; 2008) normally requires the allele frequencies in the founder population. As these frequencies are usually not known, they are commonly replaced by the observed frequencies, as we did in our study. $\mathbf{G_N}$ was derived from $\mathbf{G_{OF}}$, with normalization to provide more realistic variance and accuracy estimations (Forni et al. 2011):

$$\mathbf{G_N} = \frac{1}{2} \times \frac{(\mathbf{Z}-\mathbf{P})(\mathbf{Z}-\mathbf{P})^t}{\{\mathrm{trace}[(\mathbf{Z}-\mathbf{P})(\mathbf{Z}-\mathbf{P})^t]\}/n}$$

where n is the number of genotyped individuals.

*Multivariate models*

Multivariate models were built from [1], as follows (Mrode 2005):

$$\begin{bmatrix}\mathbf{ABW}\\ \mathbf{BN}\end{bmatrix} = \begin{bmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{X}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_{\mathrm{ABW}}\\ \boldsymbol{\beta}_{\mathrm{BN}}\end{bmatrix} + \begin{bmatrix}\mathbf{W} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}\end{bmatrix}\begin{bmatrix}\mathbf{b}_{\mathrm{ABW}}\\ \mathbf{b}_{\mathrm{BN}}\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_{\mathrm{A}} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{\mathrm{A}}\end{bmatrix}\begin{bmatrix}\mathbf{g}_{\mathbf{A_{ABW}}}\\ \mathbf{g}_{\mathbf{A_{BN}}}\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_{\mathrm{B}} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{\mathrm{B}}\end{bmatrix}\begin{bmatrix}\mathbf{g}_{\mathbf{B_{ABW}}}\\ \mathbf{g}_{\mathbf{B_{BN}}}\end{bmatrix} + \begin{bmatrix}\mathbf{Z}_{\mathrm{D}} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_{\mathrm{D}}\end{bmatrix}\begin{bmatrix}\mathbf{s}_{\mathbf{ABW}}\\ \mathbf{s}_{\mathbf{BN}}\end{bmatrix} + \begin{bmatrix}\mathbf{e}_{\mathbf{ABW}}\\ \mathbf{e}_{\mathbf{BN}}\end{bmatrix}$$

In multivariate T-BLUP, genetic effects were structured as:

$$\begin{bmatrix}\mathbf{g}_{\mathbf{A_{ABW}}}\\ \mathbf{g}_{\mathbf{A_{BN}}}\end{bmatrix} \sim N\left(0, \begin{bmatrix}\sigma^2_{g_{\mathrm{A_{ABW}}}} & \mathrm{C}_{g_{\mathrm{A}}}\\ \mathrm{C}_{g_{\mathrm{A}}} & \sigma^2_{g_{\mathrm{A_{BN}}}}\end{bmatrix} \otimes \mathbf{A_A}\right)$$

$$\begin{bmatrix}\mathbf{g}_{\mathbf{B_{ABW}}}\\ \mathbf{g}_{\mathbf{B_{BN}}}\end{bmatrix} \sim N\left(0, \begin{bmatrix}\sigma^2_{g_{\mathrm{B_{ABW}}}} & \mathrm{C}_{g_{\mathrm{B}}}\\ \mathrm{C}_{g_{\mathrm{B}}} & \sigma^2_{g_{\mathrm{B_{BN}}}}\end{bmatrix} \otimes \mathbf{A_B}\right)$$

$$\begin{bmatrix}\mathbf{s}_{\mathbf{ABW}}\\ \mathbf{s}_{\mathbf{BN}}\end{bmatrix} \sim N\left(0, \begin{bmatrix}\sigma^2_{s_{\mathrm{ABW}}} & \mathrm{C}_{\mathrm{s}}\\ \mathrm{C}_{\mathrm{s}} & \sigma^2_{s_{\mathrm{BN}}}\end{bmatrix} \otimes \mathbf{D}\right)$$

where $C_{gA}$ and $C_{gB}$ are additive genetic covariances and $C_s$ is the dominance genetic covariance. Residual effects were structured as:

$$\begin{bmatrix}\mathbf{e}_{\mathbf{ABW}}\\ \mathbf{e}_{\mathbf{BN}}\end{bmatrix} \sim N\left(0, \begin{bmatrix}\sigma^2_{e_{\mathrm{ABW}}} & \mathrm{C}_{\mathrm{e}}\\ \mathrm{C}_{\mathrm{e}} & \sigma^2_{e_{\mathrm{BN}}}\end{bmatrix} \otimes \mathbf{I}\right)$$

For multivariate G-BLUP, $\mathbf{A_A}$, $\mathbf{A_B}$ and $\mathbf{D}$ were replaced by $\mathbf{H_A}$, $\mathbf{H_B}$ and $\mathbf{D_H}$, respectively.

Non-genetic random effects had unstructured variances-covariances.

Variances and covariances of both non-genetic effects and genetic effects were estimated by REML.

**Comparison of models**

For a given type of model (i.e. univariate for ABW, univariate for BN and bivariate), the G-BLUP approaches based on the three additive genomic matrices were compared between themselves and with the T-BLUP model. At this stage, for computational reasons, the SCA effects were considered uncorrelated between traits in the multivariate models. Only the additive genetic variance-covariance structure matrix varied, while the number of observations and estimated parameters remained constant. The models were consequently directly compared based on their deviance (-2LogLikelihood), which was the equivalent of comparing their Akaike information criterion and Bayesian information criterion. The convergence of REML algorithm was also considered.

The univariate G-BLUP and multivariate G-BLUP models were compared based on their accuracy, which is the correlation between the predicted genetic effects (GCAs or SCAs) and their true value (unknown).

The accuracy of the genetic effect predicted for the x[th] level (i.e. parent for GCA or cross for SCA) was estimated from its relation with the prediction error variance (PEV) (Clark et al. 2012):

$$r_x = \sqrt{1 - \frac{PEV_x}{\sigma^2 \mathbf{\Sigma}_{xx}}}$$

where σ² is the variance of the genetic effect, $\mathbf{\Sigma}_{xx}$ is the x[th] term of the diagonal of the associated variance-covariance matrix and $PEV_x = (\hat{u} - u)^2_x$, with u the genetic effect considered. PEVs were computed from the elements of the inverse of the mixed model equations, based on theoretical derivations from Henderson (1975) (i.e. not obtained by cross-validation). Consequently, for any progeny-tested individual *x*, the accuracy associated with its GCA for a given trait was:

$$r_{GCA_x} = \sqrt{1 - \frac{PEV_{GCA_x}}{\mathbf{G}_{xx}\sigma^2_g}} \qquad [3]$$

where $\sigma^2_g$ is the estimated additive variance of the trait for the parental group of *x*. For the univariate and multivariate G-BLUP and for each trait, we computed the mean $r_{GCA}$ over the 140 group A parents and the 131 group B parents that were genotyped. For each group and each trait, the mean $r_{GCA}$ of univariate and multivariate models was compared using a t-test and a Bonferroni correction. We also compared the accuracy of SCA, which for cross *x* × *y* and a given trait, was:

$$r_{SCA_{xy}} = \sqrt{1 - \frac{PEV_{SCA_{xy}}}{\mathbf{D}_{xy\,xy}\sigma^2_s}} \qquad [4]$$

where $\mathbf{D}_{xy\,xy} = G_{xx}G_{yy}$ and $\sigma^2_s$ the estimated dominance variance for the trait. For the univariate and multivariate G-BLUP and for each trait, we computed the mean $r_{SCA}$ over the 478 crosses evaluated in the progeny test and over 256 crosses that had not been evaluated. These 256 crosses were sampled from the unevaluated crosses among all possible crosses between the 140 A and 131 B parents, with a balanced representation of the parents of both groups (i.e. each parent occurred once or twice among the 256 unevaluated crosses). To obtain the $PEV_{SCA}$ of the 256 unevaluated crosses and compute their $r_{SCA}$, these crosses were added to the **D** matrix prior to analyzing the mixed models, following Henderson (1977). The mean $r_{SCA}$ of the univariate and multivariate models was compared using a t-test for each group and trait, and a Bonferroni correction was applied to adjust the p-values.

**Marker density**

We studied the effect of marker density on the prediction of GCAs by the multivariate G-BLUP model. This was investigated independently in the two parental groups by varying the number of markers for one group, while keeping the maximum number of markers for the other group. The number of markers m varied from 10 to 265 in group A and from 10 to 289 in group B, with a step of 10. At each density, five replicates were made, for each replicate, we used a random subset of m markers chosen among all the available polymorphic markers for the group. For each replicate, the additive coancestry matrix of the group concerned was calculated using the m markers, and the dominance matrix was calculated using the m markers for the group concerned and all the markers of the other group. To assess the effect of the number of markers on the prediction of GCAs, we calculated the prediction accuracy of the model, i.e. the correlation between the predicted GCAs (for the group whose marker density varied) and the reference GCAs. The reference GCAs were obtained from the most accurate model previously identified (actually the multivariate G-BLUP) using all the markers, so that the prediction accuracy was the best approximation of accuracy. The different levels of marker number, of replicates per level of marker number and the two parental groups meant the calculations had to be repeated many times, so, to speed up the process, no covariance was specified for the dominance effects in the multivariate G-BLUP model used here.

## 3. Results

### Coancestry matrices

The distribution of coancestry estimates in group A and group B is shown in Figure 1. Coancestry estimates in $\mathbf{G_{AIS}}$ and **A** belonged to [0, 1], as expected, as coancestry is the probability that two alleles on a random locus of two individuals are identical by descent (Wright 1922; Malécot 1948). The median value of the two VanRaden matrices ($\mathbf{G_{OF}}$, $\mathbf{G_N}$) was below 0, meaning that more than half the coancestry estimates were negative. The REML algorithm converged with **A** and $\mathbf{G_{AIS}}$ matrices. The smallest deviance was obtained with $\mathbf{G_{AIS}}$ (Table 1). The $\mathbf{G_{OF}}$ and $\mathbf{G_N}$ matrices were not positive definite and the REML algorithm did not converge, leading to higher deviances than with $\mathbf{G_{AIS}}$ and **A**. Therefore, for our dataset, the $\mathbf{G_{AIS}}$ matrix appeared to be more appropriate than the other genomic matrices $\mathbf{G_{OF}}$ and $\mathbf{G_N}$, and than the genealogical matrix **A**. For the rest of the study, we consequently only used $\mathbf{G_{AIS}}$ in the G-BLUP.

For both A and B groups, coancestry estimates in $\mathbf{G_{AIS}}$ were higher than in matrix **A**. $\mathbf{G_{AIS}}$ did not contain any null coancestry estimates, whereas **A** contained 73.6% null coancestry estimates for group A and 42.9% for group B. The coancestry estimates for group A were lower than those for group B in the **A** matrix, but were higher in $\mathbf{G_{AIS}}$. The variability in coancestry estimates was higher in group A than in group B.

### Multivariate G-BLUP

The multivariate G-BLUP revealed very high additive correlations, reaching -0.997 in the parental group A and -0.917 in group B, very high dominance correlations (-0.987) and low residual correlations (-0.158).

The GCA accuracy of the univariate and multivariate G-BLUP are depicted in Figure 2A. For all combinations of groups and traits, mean GCA accuracy was higher with the multivariate G-BLUP model than with the univariate G-BLUP ($p<10^{-100}$). The average increase was 22.5%, ranging from 13.2% for ABW in group B to 32.1% for BN in group B. There were differences in GCA accuracy between traits within a parental group with univariate G-BLUP, but the multivariate G-BLUP model increased the GCA accuracy of both traits to the same level, i.e. 0.83 in group B and 0.88 in group A. Thus, the trait with the lowest GCA accuracy in the univariate models (ABW for group A and BN for group B) benefited the most from the multivariate model.

As the multivariate G-BLUP model predicted GCAs best, we used the GCAs predicted by this model as reference GCAs. The Pearson correlation coefficients between GCAs predicted by any of the models (univariate or multivariate, T-BLUP or G-BLUP) and reference GCAs are listed for each trait and each group in Table 2, as well as the Spearman's rank correlation of the 10% best individuals ("best" when evaluated by the reference

model). The GCAs obtained with univariate T-BLUP were generally the least correlated with the reference GCAs, with an average Pearson correlation coefficient of 0.946 and Spearman's correlation coefficient of 0.527. According to the Pearson correlation, the GCAs obtained with the multivariate T-BLUP and univariate G-BLUP models were highly correlated with the reference GCAs (average Pearson correlation coefficient of 0.978 and 0.966, respectively). However, the Spearman's correlation coefficients computed on the top 10% individuals were not as high, with an average value of 0.595, ranging from 0.213 to 0.978. This indicated that the model impacted the selection of the progeny tested individuals, and was therefore of importance for practical breeding. In addition, the multivariate T-BLUP gave GCAs with ranks that were more correlated with the ranks of the reference GCAs than the univariate G-BLUP (Spearman's rank correlation coefficient of 0.696 and 0.562, respectively). Therefore, the improvement obtained in the GCA estimates when using a multivariate genomic approach compared to the conventional univariate T-BLUP resulted more from the multivariate analyze than from the use of the genomic data.

SCA accuracy was higher with the multivariate G-BLUP than with the univariate G-BLUP ($p<10^{-100}$) (Figure 2B and C). The average increase in SCA accuracy was 18.7%, ranging from 12.9% (trait BN, unevaluated crosses) to 24.6% (trait ABW, evaluated crosses). With the multivariate G-BLUP model, SCA accuracies were on average 0.76 for evaluated crosses and 0.68 for unevaluated crosses.

The $h^2$ obtained with the multivariate genomic model were $h^2_{BN} = 0.53$ and $h^2_{ABW} = 0.35$ in group A, and $h^2_{BN} = 0.4$ and $h^2_{ABW} = 0.79$ in group B (see Supplementary Table S2 for the detail of variances).

**Marker density**

Figure 3 shows the effect of marker density on the prediction accuracy of GCAs with the multivariate G-BLUP for ABW. The results obtained for BN were very similar (Supplementary Fig. S1), certainly due to the high genetic correlation between ABW and BN. As marker density increased, the prediction accuracy of multivariate G-BLUP also rapidly increased before reaching a plateau slightly above the prediction accuracy of multivariate T-BLUP. To outperform the prediction accuracy of the multivariate T-BLUP model, multivariate G-BLUP required 110 markers for group A and 70 markers for group B, for both ABW and BN traits. The prediction accuracies exceeded 0.99 with 160 markers for group A for both ABW and BN, and with 80 (respectively 90) markers for group B for ABW (respectively BN).

## 4. Discussion

The general combining ability (GCA) for bunch number (BN) and average bunch weight (ABW) of progeny-tested oil palms is currently obtained with a pedigree-based univariate mixed model analysis of phenotypic data of hybrid individuals. In this study, we showed that using a multivariate model and replacing the genealogical coancestry matrices by molecular matrices of realized coancestry improved the analysis, leading to better estimated GCAs. In addition, the accuracy of the SCAs, usually neglected, reached interesting levels. We also found that this could be achieved with a reduced marker density. Indeed, the number of SSR markers that enabled G-BLUP to reach the same prediction accuracy as T-BLUP was 110 for group A and 70 for group B; while 160 markers in group A and 90 in group B were needed to achieve the maximum benefit offered by the genomic approach.

**Genomic versus genealogical coancestries**

We observed many null coancestry estimates in **A**, whereas all coancestry estimates in **G** were higher than zero. This reflected the fact that the pedigrees used to estimate the **A** matrices did not reach the base of the unrelated founders of the different populations. Consequently, the pedigree-based coancestries underestimated the real coancestries, whereas the genomic coancestries were able to capture these hidden relationships, which did not appear in the pedigree. However, as $\mathbf{G_{AIS}}$ considered identity by state and **A** identity by descent, the values in $\mathbf{G_{AIS}}$ were actually overestimated if several copies of some alleles were present in the founder populations (Eding and Meuwissen 2001). Nevertheless, the G-BLUP model using $\mathbf{G_{AIS}}$ was more appropriate for the data than the T-BLUP model, as shown by its higher likelihood.

The Deli individuals, which made up most of group A, originated from four oil palms planted in 1848 in Indonesia, while the African populations in group B can be traced back to the first half of the 20$^{th}$ century, with around 15 to 20 founders (Corley and Tinker 2003). The higher $\mathbf{G_{AIS}}$ values found in group A than in group B is not surprising, given the longer history of inbreeding, drift and artificial selection of Deli individuals. However, the coancestry estimates for group A were lower than those for group B in the **A** matrix. This resulted from the depth (number of generations) of the pedigree and from the history of the populations constituting the parental groups. In group B, the data available on the pedigrees referred roughly to the initial generation, but the longer history of the Deli population was not covered by its pedigree, which did not go back far enough in time. This explained why, according to the pedigrees, there were fewer coancestries in group A than in group B. This also explained the fact that the number of relationships hidden in the pedigree but captured by the markers was higher in group A than in group B. This increased the benefit of using the genomic models more for group A than for

group B, as shown by the bigger increase in the correlation with reference GCAs in group A when the G-BLUP model was used instead of the T-BLUP, than in group B.

**Multivariate model**

This is the first study to investigate the benefit of using multivariate genomic models for oil palm breeding. Using empirical data, we demonstrated that multivariate genomic models improved the prediction accuracy of additive effects (GCAs). In addition, we showed that in each parental group, the trait with the lowest heritability (ABW in group A and BN in group B) benefited the most from the use of a multivariate model. Both findings are in agreement with the results of previous simulations (Calus and Veerkamp 2011; Jia and Jannink 2012; Guo et al. 2014) but, in addition to the results of these studies, we showed that genomic multivariate models also increased the prediction accuracy of dominance effects (SCAs).

In the multivariate G-BLUP model, covariance between traits is considered to be identical at each marker. This could reduce the efficiency of multivariate G-BLUP relatively to a multivariate Bayesian method that would allow marker specific covariances between traits (Guo et al. 2014). An empirical comparison of these two statistical approaches with oil palm data would thus be useful.

**Density and type of molecular markers**

In the conventional pedigree-based analysis of progeny tests, the GCA of a progeny-tested individual results from the phenotypes of its progeny and the progeny of its relatives. The measure of coancestry used in this conventional approach is an expected value, as it is based on pedigree, and may thus differ from the true coancestry. The genomic approach improves this situation as it uses the realized coancestry between progeny-tested individuals. We found that even small numbers of markers (110 in group A and 70 in group B) gave GCAs similar to those obtained with a conventional pedigree-based model, which was likely a consequence of the small effective size of the parental groups of oil palm (<10) (Cros et al. 2015b). The respective history of the parental groups, with the longer history of inbreeding, drift and artificial selection in group A than group B, led to less variable realized coancestries in group A and, as we used SSR markers instead of SNPs as in the other GS studies, to fewer alleles per marker in group A. As a consequence, group A required more markers to capture the realized coancestries than group B. This difference between groups therefore resulted from their contrasted history and from the use of a multiallelic type of markers.

When all the markers were used, we achieved higher prediction accuracy than with the pedigree-based model. However, the improvement was very limited thus indicating that the phenotypic data of the hybrid progenies play a major role in the quality of the estimation of the GCAs, while the coancestry matrices used in the model play a secondary role.

In the present study, the progeny-tested individuals were genotyped using SSR markers. This type of marker is suitable for genotyping a relatively small number of individuals with a low coverage of the genome, but the practical application of GS for breeding implies large scale genotyping capabilities, at reasonable cost. Therefore, future GS studies in oil palm will likely use SNP markers, as this would reduce the cost per data point and speed up the genotyping process. Although more SNPs are needed to achieve the same GS accuracy as with SSR markers (Solberg et al. 2008), the efficiency of the current genotyping technologies ensures that SNP density will not be a limiting factor to implement GS in oil palm. Thus, two SNP arrays have been developped for this species, with 4.5 K (Ting et al. 2014) and 200K SNP (Teh 2015); while Pootakham et al. (2015) identified over 20 K SNP using the genotyping-by-sequencing approach.

**Comparison of models**

We must emphasize that, as shown by formulas [3] and [4], accuracy based on prediction error variances (PEV) cannot be used to compare models that differ in their genetic variance covariance matrices, as the estimated variances refer to a different base population. Here, the base population implicitly used with the T-BLUP model was made up of the individuals with no known parents in the pedigrees, while with the G-BLUP model, it was made up of genotyped individuals. In other words, even for methods that yield the same estimated breeding values, accuracies obtained from the PEVs are not invariant to parameterization (Stranden and Christensen 2011). Consequently, the fact that the accuracies we obtained from the PEVs for the G-BLUP models were lower than the accuracies of T-BLUP, which were around 0.90 (Cros et al. 2015b), was not meaningful. When evaluating the potential of GS to predict the GCA of individuals that have not been progeny-tested, accuracy is often estimated using a cross-validation approach, like that used in Cros et al. (2015b) for oil palm. However, in the present study, this was not possible as we were interested in the ability of GS to predict the GCA of progeny-tested individuals, and so the approach we chose was to compare T-BLUP and G-BLUP models based on their likelihood, and to trust the best model. Similarly, when considering either G-BLUP or T-BLUP, it was not possible to use likelihood to compare the univariate and multivariate versions of the model, as

386 the datasets (phenotypic observations) differed, but using PEV-based accuracy was relevant as the variance
387 covariance matrices were the same for the univariate and multivariate models.
388
389
390 **Implications for breeding**

391 The choice of the model to analyze the progeny tests impacted the practical breeding work, as it affected
392 the ranking of the evaluated individuals and therefore the set of the selected individuals. Here, we focused on BN
393 and ABW, two major traits determining oil yield, but genomic models could be used instead of the traditional
394 pedigree based models for all the traits recorded in progeny tests, i.e. bunch quality, height increment, disease
395 symptoms (Corley and Tinker 2003; Durand-Gasselin et al. 2010), annual profile of bunch production measured
396 by the Gini coefficient (Cros et al. 2013), etc. In addition, correlated traits should be analyzed jointly in a
397 multivariate model. Here we considered two traits but a higher number of correlated traits could easily be used.
398 In oil palm, several traits are known to be correlated including the number of fruits per bunch and the average
399 fruit weight, the percentage of pulp and the percentage of kernel in the fruits. As indicated by the literature, the
400 benefit of using a multivariate approach will result from the $h^2$ of the traits included in the model and from their
401 correlation. Using the same dataset, Cros et al. (2015b) showed that GS could predict the GCA of non-progeny-
402 tested individuals for some traits in group B, in particular when the candidate individuals were highly related to
403 the training set. Here, we showed that GS was also useful to predict the GCA of progeny-tested individuals and
404 the SCA of crosses. GS is therefore a highly valuable method for oil palm breeding, even with low marker
405 density.

406 Our experimental design involved a mean number of 65 hybrid individuals per cross. It would be
407 interesting to study the effect of decreasing the number of hybrid individuals per cross in the progeny tests, as we
408 would expect the G- BLUP model to be less affected than the T-BLUP, thanks to the extra information provided
409 to the G-BLUP (realized coancestries). Reducing the number of hybrid individuals per cross would also allow
410 progeny-testing more parents, thus increasing the selection intensity without increasing the cost of the progeny
411 tests. The importance of hybrid phenotypes in the prediction of GCAs also suggests that the number of markers
412 required to predict the GCAs of non-progeny-tested individuals might be higher than the number required for
413 progeny-tested individuals. However, this point requires further investigation.

414 To our knowledge, this is the first report of accuracy of SCA for oil palm crosses. It appeared to be
415 lower than the accuracy of the GCAs, with the mean accuracy of SCA of crosses that were not evaluated in the
416 fields reaching 0.68 with the multivariate model. The low proportion of dominance variance in total genetic

variance (Purba et al. 2001; Cros 2014) indicates that dominance effects are much smaller than additive effects, making the number of individuals per cross insufficient to accurately estimate SCAs. Our results question the fact that oil palm breeders only use progeny-tests to select parents with the highest GCA, without taking SCAs into consideration, or only those of the crosses that were actually tested in the trials, which represents a small proportion of possible crosses. Although the first paper dealing with BLUP methodology in oil palm dates from the 1990s (Soh 1994), many breeding companies have not yet started using BLUP for practical breeding decisions and, those that have started, did so relatively recently. Without BLUP taking coancestries into account, the analysis of the progeny-tests only provides SCA estimates for the crosses that were evaluated. When the BLUP model is provided with pedigree information, most of the dominance relationship matrix **D** contains zero elements due to the numerous null coancestries in the **A** matrices. In these conditions, the BLUP model will yield no estimates of SCA at all or only very inaccurate estimates for crosses that were not evaluated in the trials. However, as markers are more efficient than pedigrees at capturing coancestries, the accuracy of SCAs obtained with GBLUP is high enough to make selection possible, particularly with multivariate analysis of correlated traits. For these reasons, oil palm breeders should also consider SCAs when selecting among progeny-tested parents, since, although relatively small, the extra genetic gain obtained compared to selection based only on GCAs, would come at no extra cost.

## Conflicts of interest

The authors declare no conflict of interest.

## Acknowledgements

We acknowledge SOCFINDO (Indonesia) and CRAPP (Benin) for planning and carrying out the field trials with CIRAD (France) and authorizing the use of the data for this study. We thank P. Sampers and V. Pomiès for help in genotyping. This research was partly funded by a grant from PalmElit SAS.

444

445 **Tables**

446

447 **Table 1** Deviance of the mixed model according to the coancestry matrices ($\mathbf{G_{AIS}}$, $\mathbf{G_{OF}}$, $\mathbf{G_N}$ and **A**), for average
448 bunch weight (ABW), bunch number (BN) and multivariate analysis

| | $\mathbf{G_{AIS}}$ | $\mathbf{G_{OF}}$ | $\mathbf{G_N}$ | **A** |
|---|---|---|---|---|
| ABW univariate | 401,211.2 | 418,110 | 418,997 | 401,355.2 |
| BN univariate | 656,032.6 | 661,026.2 | 661,367.8 | 656,174.2 |
| Multivariate | 1,053,287.4 | 1,053,539 | 1,053,471.8 | 1,053,557.2 |

449

450

451 **Table 2** Pearson correlation (top) and Spearman's rank correlation of the top 10% individuals (bottom) between
452 predicted GCAs produced by a G-BLUP or T-BLUP, univariate or multivariate model and the reference GCAs
453 from the multivariate G-BLUP. The Pearson correlation coefficients were calculated based on the 140 group A
454 genotyped parents and the 131 group B genotyped parents. The top 10% individuals represented 14 individuals
455 in group A and 13 in group B

| Pearson correlation | | Group A | | Group B | | |
|---|---|---|---|---|---|---|
| | | ABW | BN | ABW | BN | mean |
| Multivariate | G-BLUP | 1 | 1 | 1 | 1 | 1 |
| | T-BLUP | 0.971 | 0.971 | 0.986 | 0.983 | 0.978 |
| Univariate | G-BLUP | 0.963 | 0.980 | 0.976 | 0.946 | 0.966 |
| | T-BLUP | 0.905 | 0.960 | 0.969 | 0.947 | 0.946 |

| Spearman's rank correlation on the top 10% individuals | | Group A | | Group B | | |
|---|---|---|---|---|---|---|
| | | ABW | BN | ABW | BN | mean |
| Multivariate | G-BLUP | 1 | 1 | 1 | 1 | 1 |
| | T-BLUP | 0.732 | 0.424 | 0.978 | 0.648 | 0.696 |
| Univariate | G-BLUP | 0.789 | 0.218 | 0.830 | 0.412 | 0.562 |
| | T-BLUP | 0.635 | 0.213 | 0.890 | 0.368 | 0.527 |

456

457

458

459 **References**

460 Billotte N, Marseillac N, Risterucci A-M, Adon B, Brottier P, Baurens F-C, Singh R, Herrán
461 A, Asmady H, Billot C, Amblard P, Durand-Gasselin T, Courtois B, Asmono D,
462 Cheah SC, Rohde W, Ritter E, Charrier A (2005) Microsatellite-based high density
463 linkage map in oil palm (*Elaeis guineensis* Jacq.). Theor Appl Genet 110:754–765.
464 doi: 10.1007/s00122-004-1901-8

465 Browning SR, Browning BL (2007) Rapid and Accurate Haplotype Phasing and Missing-
466 Data Inference for Whole-Genome Association Studies By Use of Localized
467 Haplotype Clustering. Am J Hum Genet 81:1084–1097. doi: 10.1086/521987

468 Calus M, Veerkamp R (2011) Accuracy of multi-trait genomic selection using different
469 methods. Genet Sel Evol 43:26.

470 Christensen O, Lund M (2010) Genomic prediction when some animals are not genotyped.
471 Genet Sel Evol 42:2.

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

475 Corley RHV, Tinker PBH (2003) Selection and breeding. In:The Oil Palm. 4th ed. Oxford:
476 Blackwell Science Ltd Blackwell Publishing. p. 133-199.

477 Cros D (2014) Etude des facteurs contrôlant l'efficacité de la sélection génomique chez le
478 palmier à huile (*Elaeis guineensis* Jacq.). Montpellier SupAgro, 124-[147] p.

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

482 Cros D, Denis M, Sánchez L, Cochard B, Flori A, Durand-Gasselin T, Nouy B, Omoré A,
483 Pomiès V, Riou V, Suryana E, Bouvet J-M (2015b) Genomic selection prediction
484 accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). Theor
485 Appl Genet 128:397–410. doi: 10.1007/s00122-014-2439-z

486 Cros D, Flori A, Nodichao L, Omoré A, Nouy B (2013) Differential response to water balance
487 and bunch load generates diversity of bunch production profiles among oil palm
488 crosses (*Elaeis guineensis*). Trop Plant Biol 6:26–36. doi: 10.1007/s12042-013-9116-2

489 Durand-Gasselin T, Blangy L, Picasso C, de Franqueville H, Breton F, Amblard P, Cochard
490 B, Louise C, Nouy B (2010) Sélection du palmier à huile pour une huile de palme
491 durable et responsabilité sociale. OCL 17:385–392.

492 Eding H, Meuwissen THE (2001) Marker-based estimates of between and within population
493 kinships for the conservation of genetic diversity. J Anim Breed Genet 118:141–159.
494 doi: 10.1046/j.1439-0388.2001.00290.x

Forni S, Aguilar I, Misztal I (2011) Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol 43:1.

Gascon J, de Berchoux C (1964) Caractéristique de la production d'Elaeis guineensis (Jacq.) de diverses origines et de leurs croisements. Application à la sélection du palmier à huile. Oléagineux 19:75–84.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0.

Grattapaglia D (2014) Breeding forest trees by genomic selection: current progress and the way forward. In: Genomics of Plant Genetic Resources, Springer Netherlands. Tuberosa R, Graner A, Frison E, pp 651–682.

Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G (2014) Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet 15:30.

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397. doi: 10.1534/genetics.107.081190

Henderson CR (1975) Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 31:423–447.

Henderson CR (1977) Best linear unbiased prediction of breeding values not in the model for records. J Dairy Sci 60:783–787. doi: 10.3168/jds.S0022-0302(77)83935-0

Isik F (2014) Genomic selection in forest tree breeding: the concept and an outlook to the future. New For 45:379–401. doi: 10.1007/s11056-014-9422-z

Jia Y, Jannink J-L (2012) Multiple Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. Genetics. doi: 10.1534/genetics.112.144246

Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and genomic information. J Dairy Sci 92:4656–4663.

Li CC, Weeks DE, Chakravarti A (1993) Similarity of DNA fingerprints due to chance and relatedness. Hum Hered 43:45–52.

Lynch M (1988) Estimation of relatedness by DNA fingerprinting. Mol Biol Evol 5:584–599.

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sunderland, MA: Sinauer associates, Inc., 980 p.

Malécot G (1948) Les mathématiques de l'hérédité. Masson & Cie, Paris, 64 p.

Meunier J, Gascon J (1972) Le schéma général d'amélioration du palmier à huile à l'IRHO. Oléagineux 27:1–12.

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

Mrode RA (2005) Linear models for the prediction of animal breeding values, 2nd edn. CABI, Oxfordshire, UK, 344 p.

Pootakham W, Jomchai N, Ruang-areerate P, Shearman JR, Sonthirod C, Sangsrakru D, Tragoonrung S, Tangphatsornruang S (2015) Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). Genomics. doi: 10.1016/j.ygeno.2015.02.002

Purba AR, Flori A, Baudouin L, Hamon S (2001) Prediction of oil palm (Elaeis guineensis, Jacq.) agronomic performances using the best linear unbiased predictor (BLUP). Theor Appl Genet 102:787–792.

R Core Team (2014) R: a language and environment for statistical computing. Vienna, Austria: the R Foundation for Statistical Computing

Soh AC (1994) Ranking parents by best linear unbiased prediction (BLUP) breeding values in oil palm. Euphytica 76:13–21.

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and densities. J Anim Sci 86:2447–2454. doi: 10.2527/jas.2007-0010

Stranden I, Christensen O (2011) Allele coding in genomic evaluation. Genet Sel Evol 43:25.

Stuber CW, Cockerham CC (1966) Gene effects and variances in hybrid populations. Genetics 54:1279–1286.

Teh C-K (2015) Genome-wide association study of oil palm mesocarp oil yield content and its application for marker selection. Plant and Animal Genomes Conference XXIII, San Diego, CA, USA.

Ting N-C, Jansen J, Mayes S, Massawe F, Sambanthamurthi R, Cheng-Li O, Chin C, Arulandoo X, Seng T-Y, Alwee S, Ithinin M, Singh R (2014) High density SNP and SSR-based genetic maps of two independent oil palm hybrids. BMC Genomics 15:309.

Tranbarger TJ, Kluabmongkol W, Sangsrakru D, Morcillo F, Tregear JW, Tragoonrung S, Billotte N (2012) SSR markers in transcripts of genes linked to post-transcriptional and transcriptional regulatory functions during vegetative and reproductive development of Elaeis guineensis. BMC Plant Biol 12:1.

Van Nocker S, Gardiner SE (2014) Breeding better cultivars, faster: applications of new technologies for the rapid deployment of superior horticultural tree crops. Horticultural Research 1.

VanRaden PM (2007) Genomic measures of relationship and inbreeding. Interbull Bull 37:33–36.

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423.

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

568 Wong CK, Bernardo R (2008) Genomewide selection in oil palm: increasing selection gain
569 per unit time and cost with small populations. Theor Appl Genet 116:815–824.

570 Wright S (1922) Coefficients of inbreeding and relationship. Amer Nat 56:330–338.

571 Zaki NM, Singh R, Rosli R, Ismail I (2012) Elaeis oleifera Genomic-SSR Markers:
572 Exploitation in Oil Palm Germplasm Diversity and Cross-Amplification in Arecaceae.
573 Int J Mol Sci 13:4069–4088. doi: 10.3390/ijms13044069

574

575

576

577 **Figure legends**

578

579 **Figure 1** Distribution of pairwise estimates of coancestry in group A (left) and group B (right) calculated from
580 pedigree data (**A**) and markers ($\mathbf{G_{AIS}}$, alike-in-state, $\mathbf{G_{OF}}$, VanRaden matrix calculated from observed frequencies
581 and $\mathbf{G_N}$, normalized VanRaden matrix)

582

583 **Figure 2** Mean accuracy of GCA and SCA predictions obtained with univariate and multivariate G-BLUP, for
584 bunch number (BN) and average bunch weight (ABW): (**A**) GCA of genotyped parents, (**B**) SCA of crosses
585 evaluated in trials and (**C**) SCA of unevaluated crosses. All G-BLUP models used the $\mathbf{G_{AIS}}$ coancestry matrix.
586 Bars indicate standard deviation (in panel A, n=140 in group A and 131 in group B, in panel B n=478 crosses
587 and in panel C n=256 crosses)

588

589 **Figure 3** Prediction accuracy of the GCAs of genotyped parents predicted with multivariate models, depending
590 on marker density, for variable ABW in groups A (left) and B (right). The solid line shows the mean prediction
591 accuracy of the multivariate G-BLUP, using $\mathbf{G_{AIS}}$. Dotted lines represent the standard deviation (n=5 replicates
592 of random samples of polymorphic markers)

593

594

Molecular Breeding, 2016, 36(1), 36:2 http://dx.doi.org/10.1007/s11032-015-0423-1

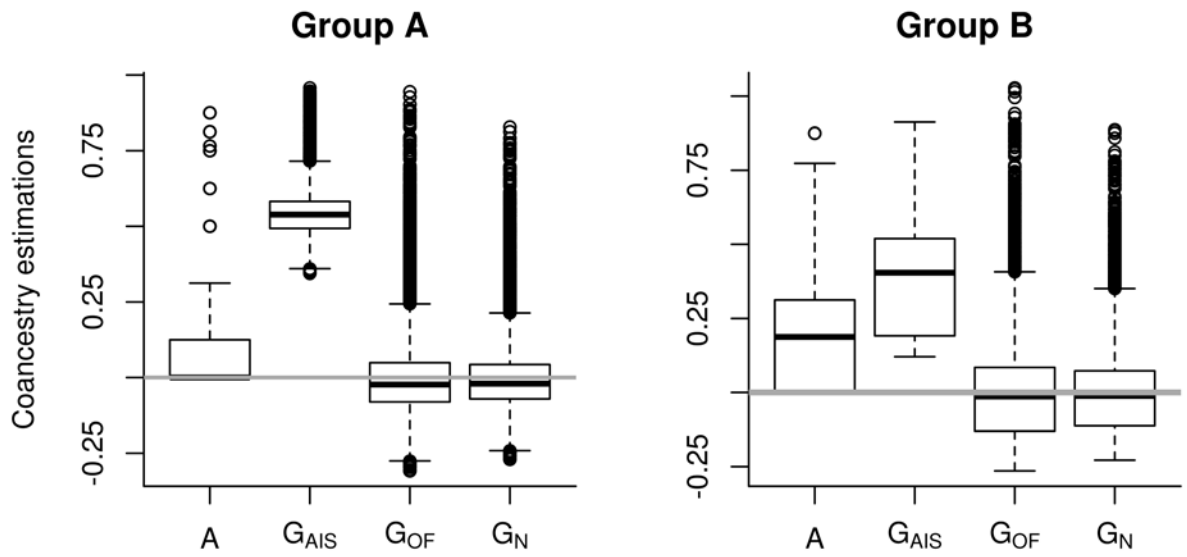


**Figure 1** Distribution of pairwise estimates of coancestry in group A (left) and group B (right) calculated from pedigree data (**A**) and markers ($\mathbf{G_{AIS}}$, alike-in-state, $\mathbf{G_{OF}}$, VanRaden matrix calculated from observed frequencies and $\mathbf{G_N}$, normalized VanRaden matrix)

Molecular Breeding, 2016, 36(1), 36:2 http://dx.doi.org/10.1007/s11032-015-0423-1

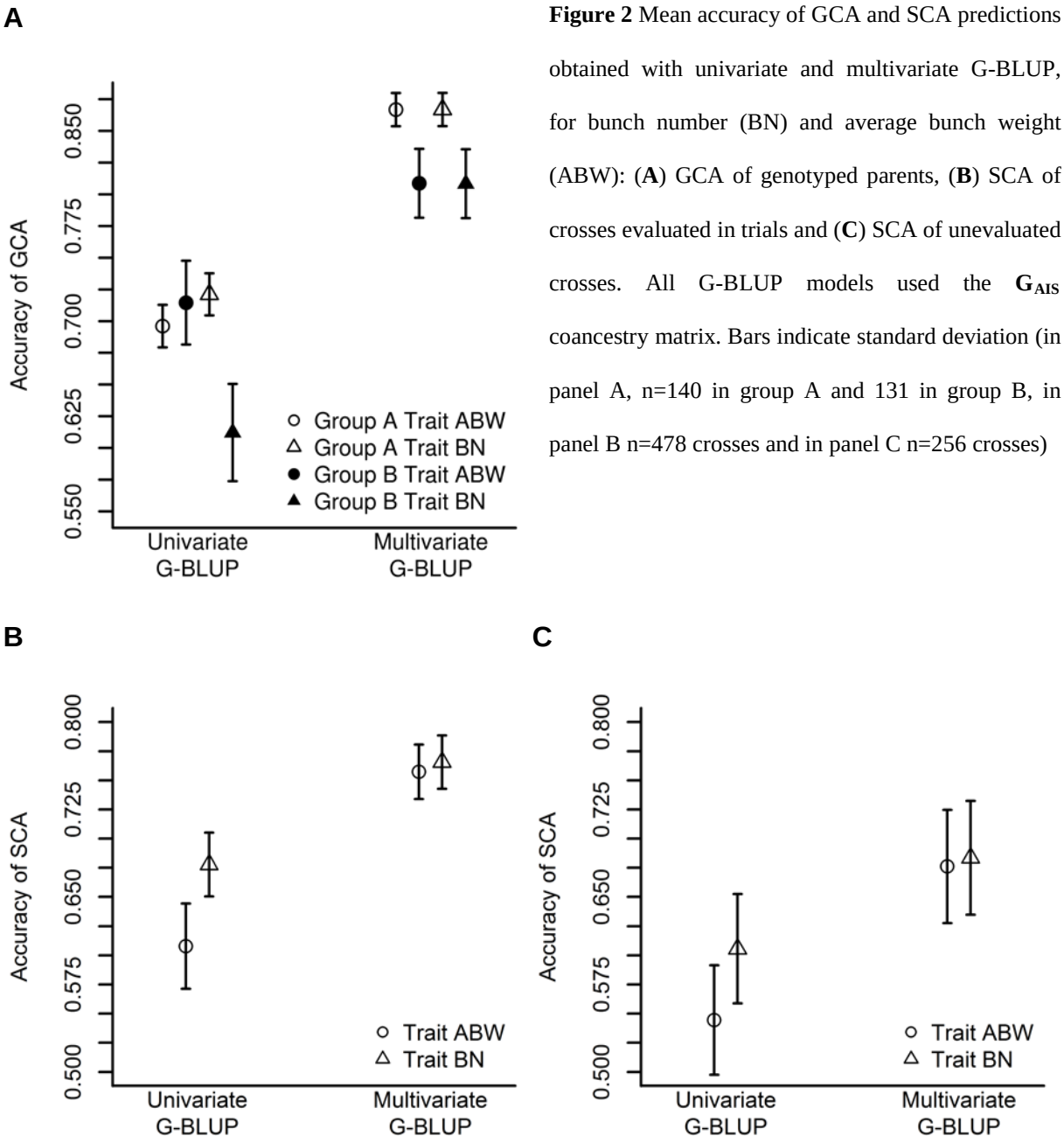


**Figure 2** Mean accuracy of GCA and SCA predictions obtained with univariate and multivariate G-BLUP, for bunch number (BN) and average bunch weight (ABW): (**A**) GCA of genotyped parents, (**B**) SCA of crosses evaluated in trials and (**C**) SCA of unevaluated crosses. All G-BLUP models used the $\mathbf{G}_{\mathbf{AIS}}$ coancestry matrix. Bars indicate standard deviation (in panel A, n=140 in group A and 131 in group B, in panel B n=478 crosses and in panel C n=256 crosses)

Molecular Breeding, 2016, 36(1), 36:2 http://dx.doi.org/10.1007/s11032-015-0423-1

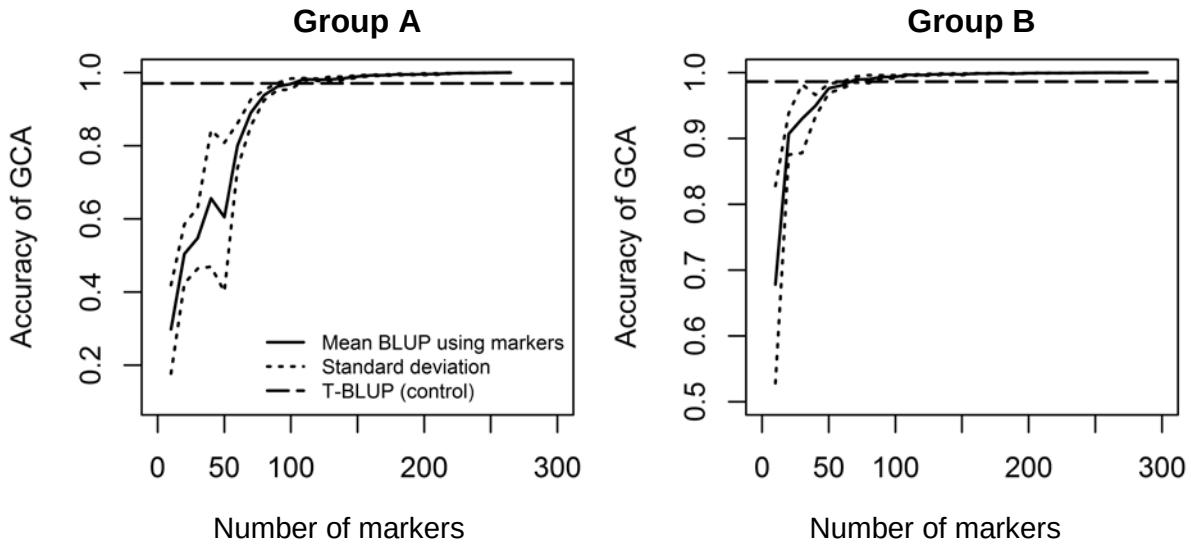


**Figure 3** Prediction accuracy of the GCAs of genotyped parents predicted with multivariate models, depending on marker density, for variable ABW in groups A (left) and B (right). The solid line shows the mean prediction accuracy of the multivariate G-BLUP, using $\mathbf{G}_{\mathbf{AIS}}$. Dotted lines represent the standard deviation (n=5 replicates of random samples of polymorphic markers)

Molecular Breeding, 2016, 36(1), 36:2 http://dx.doi.org/10.1007/s11032-015-0423-1

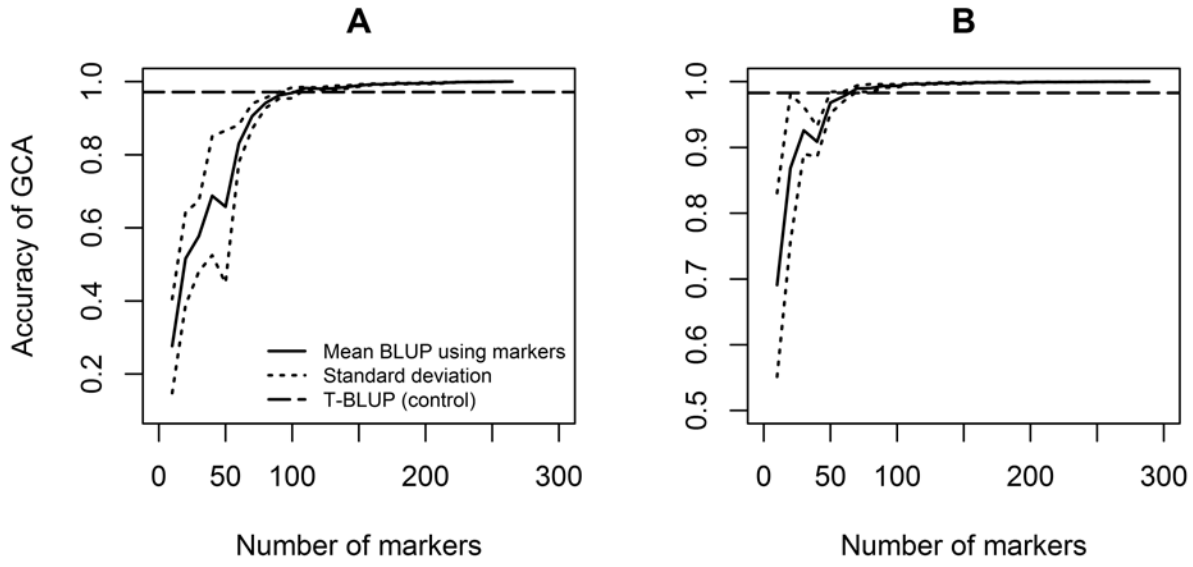


**Figure S1** Prediction accuracy of the GCAs of genotyped parents predicted with multivariate models, depending on marker density, for variable BN in groups A (panel A) and B (panel B). The solid line shows the mean prediction accuracy of the multivariate G-BLUP, using $\mathbf{G}_{\mathbf{AIS}}$. Dotted lines represent standard deviation (n=5 replicates of random samples of polymorphic markers)

Molecular Breeding, 2016, 36(1), 36:2 http://dx.doi.org/10.1007/s11032-015-0423-1

**Table S1** Details on the 271 parents used in the study, per group and population. All these individuals were present in the pedigree and genotyped.

| Group | Population | Total |
|---|---|---|
| A | | |
| | Deli | 131 |
| | Angola | 9 |
| | Total | 140 |
| B | | |
| | La Mé | 93 |
| | Yangambi | 24 |
| | Nigeria | 2 |
| | La Mé × Yangambi | 5 |
| | La Mé × Sibiti | 7 |
| | Total | 131 |

Molecular Breeding, 2016, 36(1), 36:2 http://dx.doi.org/10.1007/s11032-015-0423-1

**Table S2** Variances estimated with the multivariate genomic model for average bunch weight (ABW) and bunch number (BN): additive variances for parental groups A ($\sigma^2_{g_A}$) and B ($\sigma^2_{g_B}$) and dominance variance ($\sigma^2_s$) in A x B crosses

| | $\sigma^2_{g_A}$ | $\sigma^2_{g_B}$ | $\sigma^2_s$ |
|---|---|---|---|
| ABW | 1.15 | 2.62 | 2.81 |
| BN | 3.11 | 2.34 | 7.99 |