



HAL
open science

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.

► To cite this version:

Alexandre Marchal, Andres Legarra, Sebastien S. Tisne, Catherine Carasco-Lacombe, Aurore Manez, et al.. Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests. *Molecular Breeding*, 2016, 36 (2), Non paginé. 10.1007/s11032-015-0423-1 . hal-02640043

HAL Id: hal-02640043

<https://hal.inrae.fr/hal-02640043>

Submitted on 28 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

1 **Multivariate genomic model improves analysis of oil palm (*Elaeis* 2 *guineensis* Jacq.) progeny tests**

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

6
7 *A. Marchal, S. Tisné, C. Carasco-Lacombe, A. Manez, J.M. Bouvet, D. Cros* (✉)

8 *CIRAD, UMR AGAP (Genetic Improvement and Adaptation of Mediterranean and Tropical Plants Research*
9 *Unit), 34398 Montpellier, France*

10 *e-mail: david.cros@cirad.fr*

11
12 *Andrés Legarra*

13 *INRA, UMR1388, GenPhySe, 31326 Castanet Tolosan, France*

14
15 *E. Suryana*

16 *P.T. SOCFINDO Medan, Medan 20001, Indonesia*

17
18 *A. Omoré*

19 *INRAB, CRAPP, Pobè, Benin*

20
21 *B. Nouy, T. Durand-Gasselin*

22 *PalmElit SAS, 34980 Montferrier sur Lez, France*

23
24 *L. Sánchez*

25 *INRA, UR0588, UAGPF (Forest Tree Improvement, Genetics and Physiology Research Unit), 45075 Orléans,*
26 *France*

27
28

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

54 1. Introduction

55 Oil palm (*Elaeis guineensis* Jacq.) is the main oil crop in the world. It bears fruit bunches all year long,
56 and palm oil is extracted from the mesocarp of the fruits. Bunch production is a key component of oil yield, and
57 results from the product of two negatively correlated traits, bunch number (BN) and average bunch weight
58 (ABW) (Corley and Tinker 2003). Commercial oil palms are hybrids between two heterotic groups called A and
59 B. Group A is mostly made up of the Deli population (Asia) and group B of various African populations. Group
60 A palms have a few heavy bunches whereas group B palms have many small bunches, resulting in heterosis of
61 bunch yield in A × B hybrids. This led to the choice of a reciprocal recurrent selection (RRS) breeding scheme in
62 the 1950s (Gascon and de Berchoux 1964; Meunier and Gascon 1972). RRS involves progeny tests in which
63 group A and group B parents are crossed to estimate their general combining ability (GCA), i.e. half their
64 breeding value in hybrid crosses, for each yield component, from the phenotype of their hybrid progenies. So far,
65 parental GCAs are obtained using an univariate mixed-model analysis (i.e. considering one trait at a time) taking
66 pedigree information into account (Soh 1994; Purba et al. 2001). The accuracy of the GCAs (i.e. the correlation
67 between the estimated and the true GCAs) is high, reaching around 0.9 for all yield components (Cros et al.
68 2015b). However, the progeny tests require a long generation interval (around 20 years) and low selection
69 intensity (less than 200 individuals tested per parental group and generation).

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

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

Postprint

Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Molecular Breeding, 2016, 36(1), 36:2 <http://dx.doi.org/10.1007/s11032-015-0423-1>

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

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

94

95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125

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 ($h^2_{BN} = 0.31 \pm 0.04$ [s.e.], $h^2_{ABW} = 0.23 \pm 0.04$) and lower than h^2_{ABW} in B ($h^2_{BN} = 0.5 \pm 0.05$, $h^2_{ABW} = 0.57 \pm 0.04$) (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 \pm 0.89 (standard deviation)), and group B had 289 polymorphic SSR (mean 6.25 alleles \pm 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:

$$\mathbf{P} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{b} + \mathbf{Z}_A\mathbf{g}_A + \mathbf{Z}_B\mathbf{g}_B + \mathbf{Z}_D\mathbf{s} + \mathbf{e} \quad [1]$$

where \mathbf{P} is the vector of hybrid phenotypes (BN or ABW), \mathbf{X} and \mathbf{W} are incidence matrices of the experimental design effects, $\boldsymbol{\beta}$ and \mathbf{b} are the vectors of fixed and random effects due to the experimental design, respectively,

126 \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
127 parents A and B, respectively, \mathbf{s} is the vector of SCA of crosses, and \mathbf{e} is the vector of residual effects.

128 The random genetic effects followed the model of Stuber and Cockerham (1966), with $\mathbf{g}_A \sim N(0,$
129 $\sigma_{g_A}^2 \times \mathbf{A}_A)$, $\mathbf{g}_B \sim N(0, \sigma_{g_B}^2 \times \mathbf{A}_B)$ and $\mathbf{s} \sim N(0, \sigma_s^2 \times \mathbf{D})$, where $\sigma_{g_A}^2$ and $\sigma_{g_B}^2$ are the additive variances of the A and
130 B parents in $A \times B$ hybrid crosses, respectively, and σ_s^2 is the variance of the dominance effects in the $A \times B$
131 population. Given the hybrid nature of the crosses, the \mathbf{A} matrices contain Malécot's coefficient of coancestry f
132 (Malécot 1948), such as $\mathbf{A}_{xy} = \{f_{xy}\}$ between individuals x and y . They were built from the pedigrees with the R
133 package synbreed (Wimmer et al. 2012). The \mathbf{D} matrix is the dominance coancestry matrix between crosses,
134 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
135 (Stuber and Cockerham 1966; Lynch and Walsh 1998).

136 Fixed effects were: overall mean, "trial" (26 levels), "block" (152 levels) and "age" (6 levels). Random
137 effects associated with the experimental design were "elementary plots" (3,464 levels), "individual" (30,872
138 levels), interaction "age*cross" (" $\alpha*s$ ", 2,855 levels); with "individual" nested in "elementary plots",
139 "elementary plots" nested in "block", and "block" nested in "trial". The random experimental design effects
140 followed a normal distribution of the form $N(0, \sigma^2 \times \mathbf{I})$, where \mathbf{I} is the identity matrix and σ^2 the associated
141 variance, with the exception of $\alpha*s$ that followed $N(0, \sigma_{\alpha*s}^2 \times \mathbf{I}_{6 \times 6} \otimes \mathbf{D})$. The errors \mathbf{e} followed $N(0, \sigma_e^2 \times$
142 $\mathbf{I}_{180872 \times 180872})$, where σ_e^2 is the residual variance.

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

147 *Univariate G-BLUP*

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

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

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

154 where \mathbf{A} is the pedigree coancestry matrix including all the individuals in the group, \mathbf{A}_{22} and \mathbf{G} are the pedigree
155 and the genomic coancestry matrices, respectively, containing only the genotyped individuals. Then \mathbf{H}_D was
156 built in the same way as in equation [2]:

$$\mathbf{D}_H = \mathbf{H}_A \otimes \mathbf{H}_B$$

157
158 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 $\mathbf{s} \sim N(0, \sigma^2_s \times \mathbf{D}_H)$.

159
160 Three different genomic additive coancestry matrices \mathbf{G} were compared: \mathbf{G}_{AIS} , \mathbf{G}_{OF} and \mathbf{G}_N . \mathbf{G}_{AIS} used
161 a similarity index (Lynch 1988; Li et al. 1993) and was defined as:

$$\mathbf{G}_{AIS} = \frac{\mathbf{Z}\mathbf{Z}^t}{4L}$$

162 where \mathbf{Z} is the genotypic matrix with as many columns as alleles, with the individuals in rows, and containing in
163 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
164 index estimates coancestry from alike-in-state (AIS) alleles, and assumes that each allele was unique in the
165 founder population that generated the population under study (Eding and Meuwissen 2001).

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

$$\mathbf{G}_{OF} = \frac{(\mathbf{Z} - \mathbf{P})(\mathbf{Z} - \mathbf{P})^t}{4 \sum_{l=1}^L (1 - \sum_a p_{la}^2)}$$

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

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

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

173 where n is the number of genotyped individuals.

174

175 *Multivariate models*

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

$$\begin{bmatrix} \mathbf{ABW} \\ \mathbf{BN} \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{ABW} \\ \boldsymbol{\beta}_{BN} \end{bmatrix} + \begin{bmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{b}_{ABW} \\ \mathbf{b}_{BN} \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_A \end{bmatrix} \begin{bmatrix} \mathbf{g}_{AABW} \\ \mathbf{g}_{ABN} \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_B & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_B \end{bmatrix} \begin{bmatrix} \mathbf{g}_{BABW} \\ \mathbf{g}_{BBN} \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_D & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_D \end{bmatrix} \begin{bmatrix} \mathbf{s}_{ABW} \\ \mathbf{s}_{BN} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_{ABW} \\ \mathbf{e}_{BN} \end{bmatrix}$$

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

$$\begin{bmatrix} \mathbf{g}_{AABW} \\ \mathbf{g}_{ABN} \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma^2_{g_{AABW}} & C_{g_A} \\ C_{g_A} & \sigma^2_{g_{ABN}} \end{bmatrix} \otimes \mathbf{A}_A)$$

$$\begin{bmatrix} \mathbf{g}_{BABW} \\ \mathbf{g}_{BBN} \end{bmatrix} \sim N(0, \begin{bmatrix} \sigma^2_{g_{BABW}} & C_{g_B} \\ C_{g_B} & \sigma^2_{g_{BBN}} \end{bmatrix} \otimes \mathbf{A}_B)$$

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

178 where C_{g_A} and C_{g_B} are additive genetic covariances and C_s is the dominance genetic covariance. Residual effects
179 were structured as:

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

180

181 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.

182 Non-genetic random effects had unstructured variances-covariances.

183

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

185

186 Comparison of models

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

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

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

$$r_x = \sqrt{1 - \frac{\text{PEV}_x}{\sigma^2 \Sigma_{xx}}}$$

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

$$r_{\text{GCA}_x} = \sqrt{1 - \frac{\text{PEV}_{\text{GCA}_x}}{G_{xx}\sigma_g^2}} \quad [3]$$

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

$$r_{\text{SCA}_{xy}} = \sqrt{1 - \frac{\text{PEV}_{\text{SCA}_{xy}}}{D_{xy \ xy} \sigma_s^2}} \quad [4]$$

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

221 Marker density

222 We studied the effect of marker density on the prediction of GCAs by the multivariate G-BLUP model.
223 This was investigated independently in the two parental groups by varying the number of markers for one group,
224 while keeping the maximum number of markers for the other group. The number of markers m varied from 10 to
225 265 in group A and from 10 to 289 in group B, with a step of 10. At each density, five replicates were made, for
226 each replicate, we used a random subset of m markers chosen among all the available polymorphic markers for
227 the group. For each replicate, the additive coancestry matrix of the group concerned was calculated using the m
228 markers, and the dominance matrix was calculated using the m markers for the group concerned and all the
229 markers of the other group. To assess the effect of the number of markers on the prediction of GCAs, we
230 calculated the prediction accuracy of the model, i.e. the correlation between the predicted GCAs (for the group
231 whose marker density varied) and the reference GCAs. The reference GCAs were obtained from the most
232 accurate model previously identified (actually the multivariate G-BLUP) using all the markers, so that the
233 prediction accuracy was the best approximation of accuracy. The different levels of marker number, of replicates
234 per level of marker number and the two parental groups meant the calculations had to be repeated many times,
235 so, to speed up the process, no covariance was specified for the dominance effects in the multivariate G-BLUP
236 model used here.

237

3. Results

Coancestry matrices

The distribution of coancestry estimates in group A and group B is shown in Figure 1. Coancestry estimates in 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 (G_{OF} , G_N) was below 0, meaning that more than half the coancestry estimates were negative. The REML algorithm converged with A and G_{AIS} matrices. The smallest deviance was obtained with G_{AIS} (Table 1). The G_{OF} and G_N matrices were not positive definite and the REML algorithm did not converge, leading to higher deviances than with G_{AIS} and A . Therefore, for our dataset, the G_{AIS} matrix appeared to be more appropriate than the other genomic matrices G_{OF} and G_N , and than the genealogical matrix A . For the rest of the study, we consequently only used G_{AIS} in the G-BLUP.

For both A and B groups, coancestry estimates in G_{AIS} were higher than in matrix A . 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 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

268 model). The GCAs obtained with univariate T-BLUP were generally the least correlated with the reference
269 GCAs, with an average Pearson correlation coefficient of 0.946 and Spearman's correlation coefficient of 0.527.
270 According to the Pearson correlation, the GCAs obtained with the multivariate T-BLUP and univariate G-BLUP
271 models were highly correlated with the reference GCAs (average Pearson correlation coefficient of 0.978 and
272 0.966, respectively). However, the Spearman's correlation coefficients computed on the top 10% individuals
273 were not as high, with an average value of 0.595, ranging from 0.213 to 0.978. This indicated that the model
274 impacted the selection of the progeny tested individuals, and was therefore of importance for practical breeding.
275 In addition, the multivariate T-BLUP gave GCAs with ranks that were more correlated with the ranks of the
276 reference GCAs than the univariate G-BLUP (Spearman's rank correlation coefficient of 0.696 and 0.562,
277 respectively). Therefore, the improvement obtained in the GCA estimates when using a multivariate genomic
278 approach compared to the conventional univariate T-BLUP resulted more from the multivariate analyze than
279 from the use of the genomic data.

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

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

286

287 **Marker density**

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

296

297 **4. Discussion**

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

307

308 **Genomic versus genealogical coancestries**

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

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

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

331 **Multivariate model**

332 This is the first study to investigate the benefit of using multivariate genomic models for oil palm
333 breeding. Using empirical data, we demonstrated that multivariate genomic models improved the prediction
334 accuracy of additive effects (GCAs). In addition, we showed that in each parental group, the trait with the lowest
335 heritability (ABW in group A and BN in group B) benefited the most from the use of a multivariate model. Both
336 findings are in agreement with the results of previous simulations (Calus and Veerkamp 2011; Jia and Jannink
337 2012; Guo et al. 2014) but, in addition to the results of these studies, we showed that genomic multivariate
338 models also increased the prediction accuracy of dominance effects (SCAs).

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

344 **Density and type of molecular markers**

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

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

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

371 **Comparison of models**

372 We must emphasize that, as shown by formulas [3] and [4], accuracy based on prediction error
373 variances (PEV) cannot be used to compare models that differ in their genetic variance covariance matrices, as
374 the estimated variances refer to a different base population. Here, the base population implicitly used with the T-
375 BLUP model was made up of the individuals with no known parents in the pedigrees, while with the G-BLUP
376 model, it was made up of genotyped individuals. In other words, even for methods that yield the same estimated
377 breeding values, accuracies obtained from the PEVs are not invariant to parameterization (Stranden and
378 Christensen 2011). Consequently, the fact that the accuracies we obtained from the PEVs for the G-BLUP
379 models were lower than the accuracies of T-BLUP, which were around 0.90 (Cros et al. 2015b), was not
380 meaningful. When evaluating the potential of GS to predict the GCA of individuals that have not been progeny-
381 tested, accuracy is often estimated using a cross-validation approach, like that used in Cros et al. (2015b) for oil
382 palm. However, in the present study, this was not possible as we were interested in the ability of GS to predict
383 the GCA of progeny-tested individuals, and so the approach we chose was to compare T-BLUP and G-BLUP
384 models based on their likelihood, and to trust the best model. Similarly, when considering either G-BLUP or T-
385 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.

389 **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

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

435 Conflicts of interest

436 The authors declare no conflict of interest.

438 Acknowledgements

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

Tables

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

	G_{AIS}	G_{OF}	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

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

Pearson correlation		Group A		Group B		mean
		ABW	BN	ABW	BN	
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		mean
		ABW	BN	ABW	BN	
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

References

- 458
459
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-Gasselín 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 make-up 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-Gasselín 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-Gasselín 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

- 495 Forni S, Aguilar I, Misztal I (2011) Different genomic relationship matrices for single-step
496 analysis using phenotypic, pedigree and genomic information. *Genet Sel Evol* 43:1.
- 497 Gascon J, de Berchoux C (1964) Caractéristique de la production d'*Elaeis guineensis* (Jacq.)
498 de diverses origines et de leurs croisements. Application à la sélection du palmier à
499 huile. *Oléagineux* 19:75–84.
- 500 Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0.
- 501 Grattapaglia D (2014) Breeding forest trees by genomic selection: current progress and the
502 way forward. In: *Genomics of Plant Genetic Resources*, Springer Netherlands.
503 Tuberosa R, Graner A, Frison E, pp 651–682.
- 504 Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G (2014) Comparison of single-trait and
505 multiple-trait genomic prediction models. *BMC Genet* 15:30.
- 506 Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information
507 on genome-assisted breeding values. *Genetics* 177:2389–2397. doi:
508 10.1534/genetics.107.081190
- 509 Henderson CR (1975) Best Linear Unbiased Estimation and Prediction under a Selection
510 Model. *Biometrics* 31:423–447.
- 511 Henderson CR (1977) Best linear unbiased prediction of breeding values not in the model for
512 records. *J Dairy Sci* 60:783–787. doi: 10.3168/jds.S0022-0302(77)83935-0
- 513 Isik F (2014) Genomic selection in forest tree breeding: the concept and an outlook to the
514 future. *New For* 45:379–401. doi: 10.1007/s11056-014-9422-z
- 515 Jia Y, Jannink J-L (2012) Multiple Trait Genomic Selection Methods Increase Genetic Value
516 Prediction Accuracy. *Genetics*. doi: 10.1534/genetics.112.144246
- 517 Legarra A, Aguilar I, Misztal I (2009) A relationship matrix including full pedigree and
518 genomic information. *J Dairy Sci* 92:4656–4663.
- 519 Li CC, Weeks DE, Chakravarti A (1993) Similarity of DNA fingerprints due to chance and
520 relatedness. *Hum Hered* 43:45–52.
- 521 Lynch M (1988) Estimation of relatedness by DNA fingerprinting. *Mol Biol Evol* 5:584–599.
- 522 Lynch M, Walsh B (1998) *Genetics and analysis of quantitative traits*. Sunderland, MA:
523 Sinauer associates, Inc., 980 p.
- 524 Malécot G (1948) *Les mathématiques de l'hérédité*. Masson & Cie, Paris, 64 p.
- 525 Meunier J, Gascon J (1972) Le schéma général d'amélioration du palmier à huile à l'IRHO.
526 *Oléagineux* 27:1–12.
- 527 Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using
528 genome-wide dense marker maps. *Genetics* 157:1819–1829.

- 529 Mrode RA (2005) Linear models for the prediction of animal breeding values, 2nd edn.
530 CABI, Oxfordshire, UK, 344 p.
- 531 Pootakham W, Jomchai N, Ruang-areerate P, Shearman JR, Sonthirod C, Sangsrakru D,
532 Tragoonrung S, Tangphatsornruang S (2015) Genome-wide SNP discovery and
533 identification of QTL associated with agronomic traits in oil palm using genotyping-
534 by-sequencing (GBS). *Genomics*. doi: 10.1016/j.ygeno.2015.02.002
- 535 Purba AR, Flori A, Baudouin L, Hamon S (2001) Prediction of oil palm (*Elaeis guineensis*,
536 Jacq.) agronomic performances using the best linear unbiased predictor (BLUP).
537 *Theor Appl Genet* 102:787–792.
- 538 R Core Team (2014) R: a language and environment for statistical computing. Vienna,
539 Austria: the R Foundation for Statistical Computing
- 540 Soh AC (1994) Ranking parents by best linear unbiased prediction (BLUP) breeding values in
541 oil palm. *Euphytica* 76:13–21.
- 542 Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using
543 different marker types and densities. *J Anim Sci* 86:2447–2454. doi:
544 10.2527/jas.2007-0010
- 545 Strandén I, Christensen O (2011) Allele coding in genomic evaluation. *Genet Sel Evol* 43:25.
- 546 Stuber CW, Cockerham CC (1966) Gene effects and variances in hybrid populations.
547 *Genetics* 54:1279–1286.
- 548 Teh C-K (2015) Genome-wide association study of oil palm mesocarp oil yield content and
549 its application for marker selection. Plant and Animal Genomes Conference XXIII,
550 San Diego, CA, USA.
- 551 Ting N-C, Jansen J, Mayes S, Massawe F, Sambanthamurthi R, Cheng-Li O, Chin C,
552 Arulandoo X, Seng T-Y, Alwee S, Ithinin M, Singh R (2014) High density SNP and
553 SSR-based genetic maps of two independent oil palm hybrids. *BMC Genomics*
554 15:309.
- 555 Tranbarger TJ, Kluabmongkol W, Sangsrakru D, Morcillo F, Tregear JW, Tragoonrung S,
556 Billotte N (2012) SSR markers in transcripts of genes linked to post-transcriptional
557 and transcriptional regulatory functions during vegetative and reproductive
558 development of *Elaeis guineensis*. *BMC Plant Biol* 12:1.
- 559 Van Nocker S, Gardiner SE (2014) Breeding better cultivars, faster: applications of new
560 technologies for the rapid deployment of superior horticultural tree crops.
561 *Horticultural Research* 1.
- 562 VanRaden PM (2007) Genomic measures of relationship and inbreeding. *Interbull Bull*
563 37:33–36.
- 564 VanRaden PM (2008) Efficient methods to compute genomic predictions. *J Dairy Sci*
565 91:4414–4423.

Postprint

Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Molecular Breeding, 2016, 36(1), 36:2 <http://dx.doi.org/10.1007/s11032-015-0423-1>

- 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

Manuscrit d'auteur / Author Manuscript

Manuscrit d'auteur / Author Manuscript

Manuscrit d'auteur / Author Manuscript

Figure legends

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

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 **G**_{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)

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 **G**_{AIS}. Dotted lines represent the standard deviation (n=5 replicates of random samples of polymorphic markers)

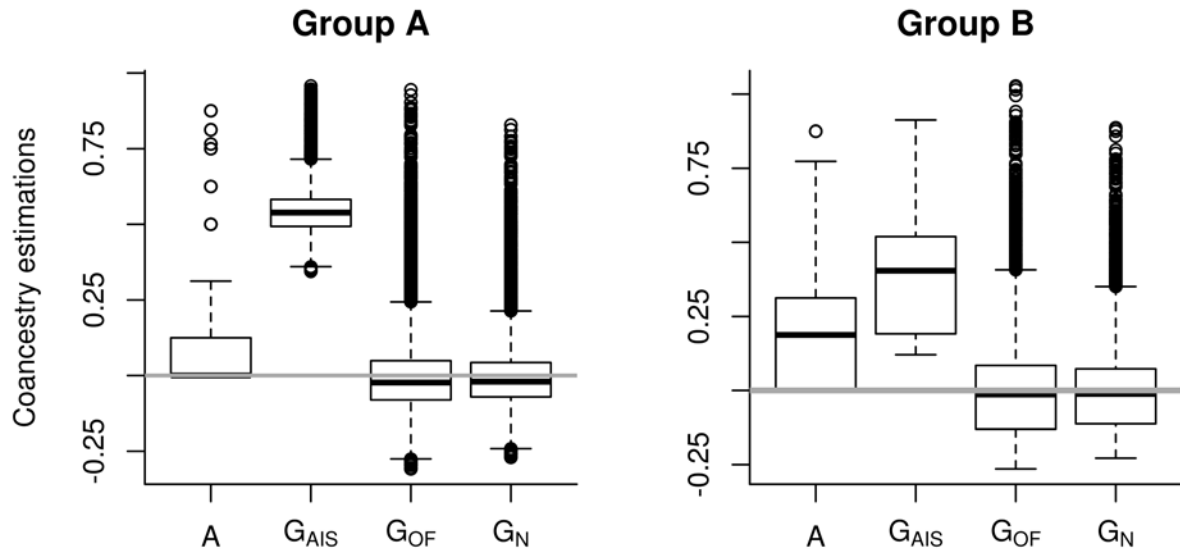
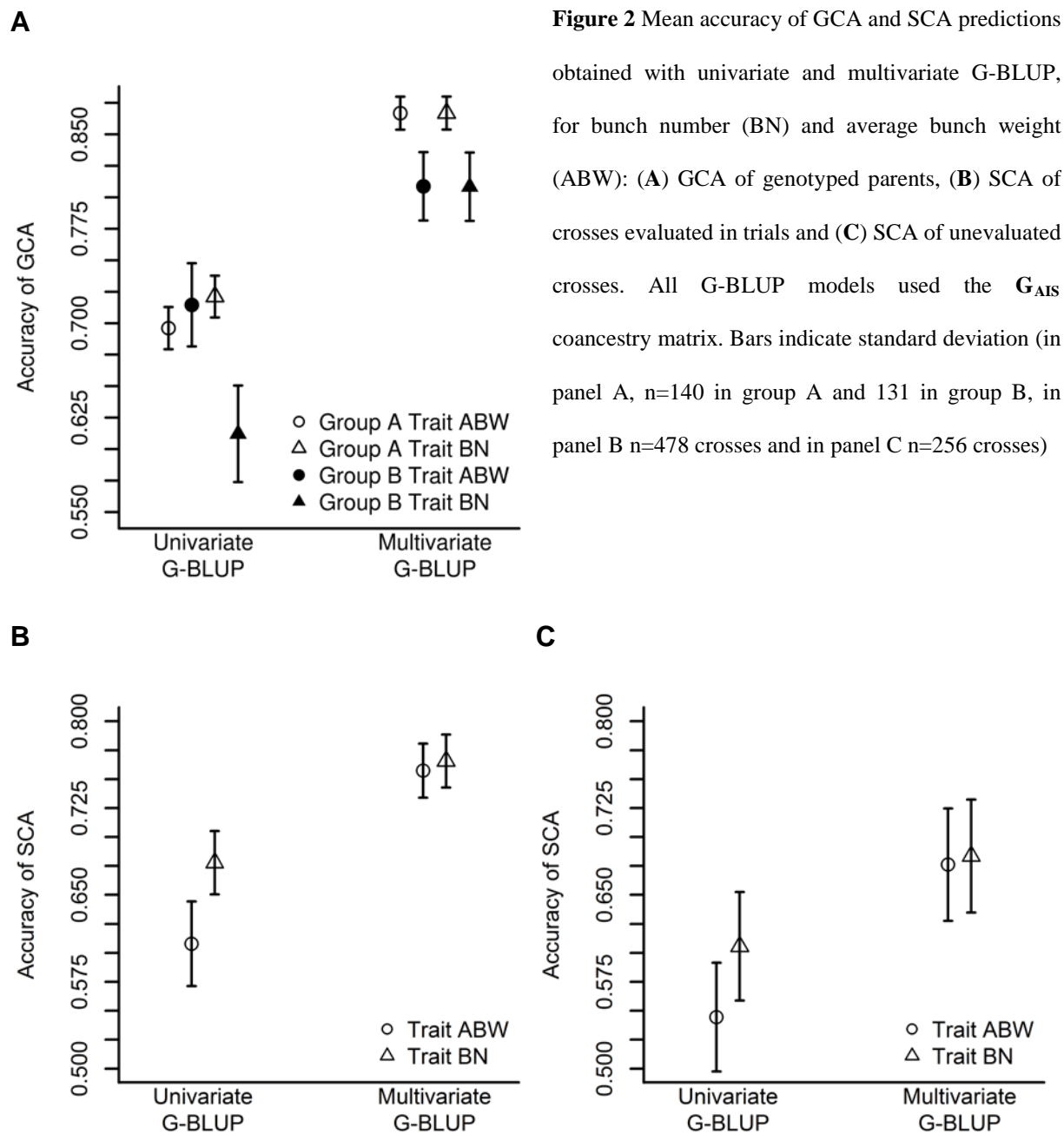


Figure 1 Distribution of pairwise estimates of coancestry in group A (left) and group B (right) calculated from pedigree data (**A**) and markers (**G_{AIS}**, alike-in-state, **G_{OF}**, VanRaden matrix calculated from observed frequencies and **G_N**, normalized VanRaden matrix)



Comment citer ce document :

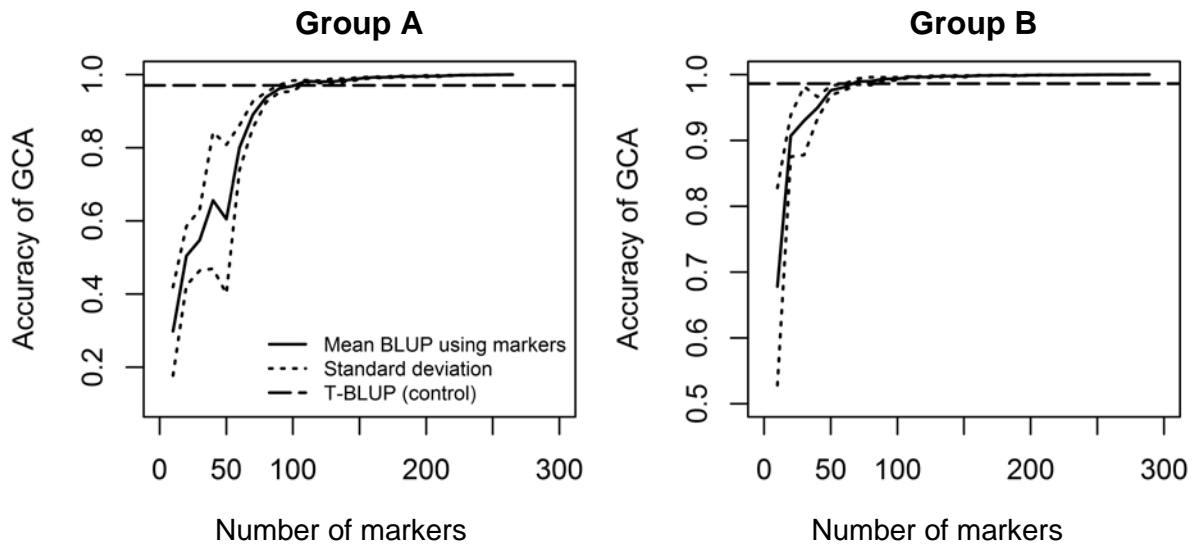


Figure 3 Prediction accuracy of the GCAs of genotyped parents predicted with multivariate models, depending on marker density, for variable ABW in groups A (left) and B (right). The solid line shows the mean prediction accuracy of the multivariate G-BLUP, using G_{AIS} . Dotted lines represent the standard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

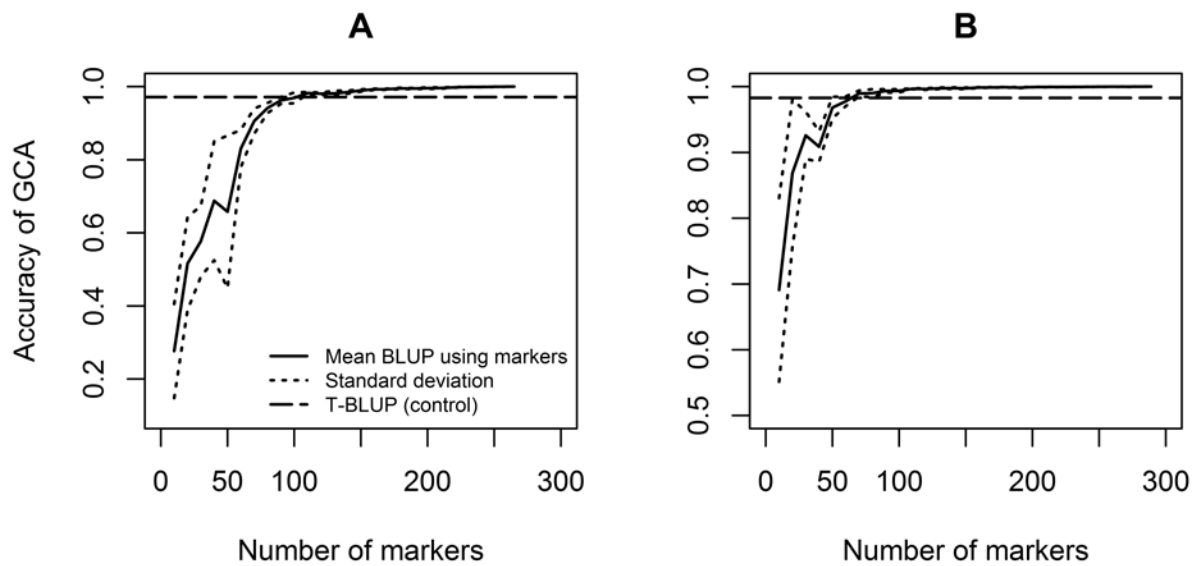


Figure S1 Prediction accuracy of the GCAs of genotyped parents predicted with multivariate models, depending on marker density, for variable BN in groups A (panel A) and B (panel B). The solid line shows the mean prediction accuracy of the multivariate G-BLUP, using \mathbf{G}_{AIS} . Dotted lines represent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

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

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 (σ^2_s) in A x B crosses

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