Genomic predictions improve clonal selection in oil palm (Elaeis guineensis Jacq.) hybrids

Achille Nyouma, Joseph Martin Bell, Florence Jacob, Virginie Riou, Aurore Manez, Virginie Pomiès, Leifi Nodichao, Indra Syahputra, Dadang Affandi, Benoit Cochard, et al.

1 Title

2 **Genomic predictions improve clonal selection in oil palm (*Elaeis guineensis* Jacq.) hybrids**

3 **Authors and affiliations:**

4 Achille NYOUMA[a,b], Joseph Martin BELL[a], Florence JACOB[c], Virginie RIOU[d,e], Aurore MANEZ[d,e],

5 Virginie POMIÈS[d,e], Leifi NODICHAO[e,f], Indra SYAHPUTRA[g], Dadang AFFANDI[g], Benoit

6 COCHARD[c], Tristan DURAND-GASSELIN[c], David CROS[b,d,e*]

7 [a] Department of Plant Biology, Faculty of Science, University of Yaoundé I, Yaoundé, Cameroon

8 [b] CETIC (African Center of Excellence in Information and Communication Technologies), University

9 of Yaoundé 1, Yaoundé, Cameroon

10 [c] PalmElit SAS, 34980 Montferrier sur Lez, France

11 [d] CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement),

12 UMR AGAP, F-34398 Montpellier, France.

13 [e] University of Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

14 [f] INRAB, CRA-PP, Pobè, Benin

15 [g] P.T. SOCFINDO Medan, Medan, Indonesia

16 [*] Corresponding author.

17 E-mail address: david.cros@cirad.fr (D. Cros).

18

19

20

21

22

Abstract

The prediction of clonal genetic value for yield is challenging in oil palm (*Elaeis guineensis* Jacq.). Currently, clonal selection involves two stages of phenotypic selection (PS): ortet preselection on traits with sufficient heritability among a small number of individuals in the best crosses in progeny tests, and final selection on performance in clonal trials. The present study evaluated the efficiency of genomic selection (GS) for clonal selection. The training set comprised almost 300 Deli × La Mé crosses phenotyped for eight palm oil yield components and the validation set 42 Deli × La Mé ortets. Genotyping-by-sequencing (GBS) revealed 15,054 single nucleotide polymorphisms (SNP). The effects of the SNP dataset (density and percentage of missing data) and two GS modeling approaches, ignoring (ASGM) and considering (PSAM) the parental origin of alleles, were assessed. The results showed prediction accuracies ranging from -0.03 to 0.70 for ortet candidates without data records, depending on trait, SNP dataset and modeling. ASGM was better (more robust over traits and SNP datasets, and simpler), although PSAM could slightly improve prediction accuracies for the two traits defining the heterotic groups. The number of SNPs had to reach 7,000, while the percentage of missing data per SNP was of secondary importance. GS prediction accuracies were higher than those of PS for most of the traits. Finally, this makes possible two practical applications of GS, that will increase genetic progress by improving ortet preselection before clonal trials: (1) preselection at the mature stage on all yield components jointly using ortet genotypes and phenotypes, and (2) genomic preselection on more yield components than PS, among a large population of the best possible crosses at nursery stage.

Keywords *Elaeis guineensis*, genomic selection, ortets, clonal selection, genotyping-by-sequencing, prediction accuracy

1. Introduction

The annual yield of palm oil is around four tons per hectare and world production is currently above 75 million tons of crude palm oil [1]. Most cultivated oil palms (*Elaeis guineensis* Jacq.) are hybrid cultivars, mainly due to their high yield per hectare. Two parental and heterotic groups are involved in the production of hybrid cultivars, namely group A, consisting essentially of the Deli population (Asia) and, to a lesser extent, the Angola population, and group B, involving the other African breeding populations. Group A produces a small number of large bunches and group B produces a lot of small bunches. This complementarity and the resulting heterosis expressed on hybrids through sexual crosses explains why they were widely adopted in the 1960s, leading to a 30% yield increase [2]. In addition, commercial oil palm material is of *tenera* (*T*) (thin-shelled) fruit type, resulting from the cross between the thick-shelled *dura* (*D*) of group A and the shell-less and usually female sterile *pisifera* (*P*) of group B. Selection of hybrids is carried out through progeny tests in a modified reciprocal recurrent selection (MRRS) breeding scheme [3,4]. The best hybrids are primarily selected based on the parental general combining abilities (GCA). Although the annual increase of the oil palm hybrids' yield obtained through genetic improvement reached 1-1.5% over the past decades [5], this remains insufficient to face the expected increase in the demand.

An additional yield increase of 20-30% compared to sexual crosses can be obtained by using clones (ramets) obtained from the micropropagation of top-ranking commercial hybrid *T* individuals (ortets) [6]. This allows taking advantage of the within hybrid crosses variability that results from parental heterozygosity. However, this approach has been hampered for a long time by a floral epigenetic abnormality producing mantled fruits, which could result in severe production loss. This abnormality is a somaclonal variation arising during tissue culture due to hypomethylation of the retrotransposon *Karma* in mantled variants, leading to homeotic transformations and parthenocarpy [7–9]. The recent understanding of the molecular mechanism involved in the mantled disorder has led to the possibility of early detection of mantled ramets during the first stages of seedling growth [8], thus arousing a new impetus for oil palm clonal selection. The evaluation of ortets on their phenotypic value is possible, but some of the oil palm yield components have a low heritability (e.g. [10] found a

broad-sense heritability (H^2) of 0 and 0.1 for bunch number and total bunch production, respectively), the estimation of their genetic values is thus of low reliability. As a consequence, breeders set clonal trials where they evaluate samples of ramets of candidate ortets that are preselected on the few yield traits with high heritability, i.e. usually the percentage of pulp per fruit (PF) and of oil per pulp (OP), for which, e.g., Nouy et al. [10] found H^2 values of 0.84 and 0.63, respectively. These trials give accurate estimations of the genetic value of the ortets but also extend, by around 10 years, the time required for the selection process for clone production, setting of trials and collection of phenotypic data. This considerably reduces the interest of clonal selection as, during this time, conventional hybrids were also improved. Another drawback of the clonal trials is that their cost means that only a small number of ortet candidates can be evaluated, thus limiting the selection intensity. There is, therefore, a need to optimize clonal selection in the oil palm.

Genomic selection (GS) [11] is a marker-assisted selection (MAS) method with a high density of markers on the entire genome, so that at least one marker can be in linkage disequilibrium with each quantitative trait locus (QTL) [12]. Compared to the previous MAS approach based on QTL detection, GS takes into account all the markers jointly and without any test of significance. In this way, even markers capturing small QTL effects are used in the model predicting the genetic values, thus improving the efficiency of selection. GS is, therefore, the most appropriate MAS method for yield traits which are usually quantitative, i.e. controlled by many loci with small effect. The GS model is calibrated (or trained) on individuals genotyped and phenotyped (training set), and predicts the genetic value of a set of related individuals that are genotyped with the same markers. Before its practical application, the GS method must be evaluated and the prediction model that gives the highest accuracy (i.e. the correlation between the predicted and the true genetic values) is retained [13]. The GS accuracy is estimated in a validation set, made of individuals genotyped and phenotyped and representative of the population that will be used for application. Oil palm is one of the pioneer perennial crops on which GS studies have been carried out. The oil palm GS studies provided prominent results, such as the superiority of GS over both QTL-based MAS and phenotypic selection [14], and the possibility of increasing the performance of sexual hybrid crosses by genomic preselection before progeny-tests [15]. The main advantages of GS for the oil palm are its ability to

enhance selection intensity and/or to shorten the generation interval, thus increasing the annual genetic gain [16]. A recent study using a large training set estimated the GS accuracy when predicting the phenotypes of hybrid individuals [17]. Phenotypes are estimates of the total genetic values but they often have low reliability, and therefore, when evaluating GS for clonal selection, it would be better to use clonal values as the target values predicted by the GS models. This has not yet been done in the oil palm, although the potential benefits of genomic clonal selection have already been shown in other perennial crops such as the eucalyptus [18] and the rubber tree [19].

Given that ortets come from a cross between two oil palm origins, the genomic prediction of their genetic values can be done by two modeling approaches [20], which are the genomic extensions of the modeling approach developed by Stuber and Cockerham [21] for interpopulation hybrids. The first one, the population-specific effects of single nucleotide polymorphism (SNP) alleles model (PSAM, or BSAM in the animal breeding literature, for breed instead of population), considers that alleles of the same marker have different effects in the hybrids depending on their population of origin, whereas the second approach, the across-population SNP genotype model (ASGM), considers that alleles of a marker have the same effect regardless of their population of origin. Studies in livestock showed that BSAM can outperform ASGM in terms of accuracy with a low number of SNPs, a large training set and slightly related or unrelated individuals [20]. However, to our knowledge, in the context of plant hybrids, these types of models were only compared in simulated maize populations [22].

The goals of this empirical study were: (1) to evaluate the efficiency of GS for clonal selection, using ortets of known clonal value to validate genomic predictions, (2) to compare ASGM and PSAM approaches, and (3) to evaluate the possibility of using GS instead of the current phenotypic selection to select the hybrid individuals to test in the clonal trials. The training set was composed of almost 300 Deli × La Mé crosses and the validation set of 42 Deli × La Mé ortets. The parents of the training crosses and the validation ortets were genotyped using genotyping-by-sequencing (GBS). Predictions were made for eight yield components, with three bunch production traits, i.e. bunch number (BN), average bunch weight (ABW) and total bunch production (FFB, for fresh fruit bunch), and five bunch quality traits, i.e. average fruit weight (AFW), fruit to bunch (FB),

pulp to fruit (PF) and oil to pulp (OP) ratios and number of fruits per bunch (NF). The effect of the SNP dataset (SNP density and percentage of missing data) was studied by filtering SNPs with different maximum percentages of missing data.

2. Materials and methods

2.1. *Plant materials and experimental designs*

The plant material used to train the GS model comes from controlled crosses between Deli and La Mé (LM) individuals. Deli material comes from four ancestors of an unknown area of Africa planted in Indonesia in 1848. The La Mé material used here comes from three founders collected in Ivory Coast between 1924 and 1930 [15,23]. For bunch production predictions, the training set was composed of 295 progeny-test crosses planted from 1995 to 2000 at Aek Loba Timur (ALT) and involving 108 Deli and 102 La Mé. For bunch quality predictions, a sample of 279 crosses involving 103 Deli and 100 La Mé parents were used (Table 1). The pedigrees of these populations are known over several generations (see Cros et al. [12]). ALT is located at 2° 39′ N – 99° 42′ E in North Sumatra, on the SOCFINDO estate (Indonesia) and is constituted of 28 trials planted on deep loamy sand soils, with low water deficit and high insolation, and benefiting from standard cultural practices [24]. The experimental design used in these trials was either a balanced lattice of four to five ranks or randomized complete block designs (RCBD), described in detail by Cros et al. [15].

The validation set was composed of 42 Deli × La Mé *tenera* ortets, evaluated in clonal trials involving on average 69 ramets per clone for production traits and a subset of 34 ramets per clone for quality traits. The ramets were established in three out of the 28 trials of ALT and were planted in 1995 and 1998 (Table 1). The 42 ortets were chosen among individuals from various hybrid crosses planted on seven trials of an earlier set of progeny tests, located at Aek Kwasan 1 (AK1), which was also located on the SOCFINDO estate and benefited from the same agricultural practices. The plantation of the seven trials of AK1 took place between 1975 and 1979. The 42 ortets come from 17 families of full sibs with 16 La Mé parents and 12 Deli parents. These families were composed of one to five ortets each, with four families having five ortets each.

155

156 2.2. *Phenotyping*

157 All the individuals, i.e. the training hybrid crosses, the 42 hybrid ortets and their ramets, were
158 phenotyped for eight traits. Five traits were assessed for bunch quality: average fruit weight (AFW),
159 fruit to bunch (FB), pulp to fruit (PF), and oil to pulp (OP) ratios, and number of fruits per bunch
160 (NF); and three traits for bunch production: bunch number (BN), average bunch weight (ABW), and
161 total bunch production (FFB). For quality traits, data were collected when plants were from five to
162 nine years old at ALT and from six to nine years old at AK1. For production traits, data were collected
163 when the plants were from three to seven years old in both sites.

164 The coefficients of variation (CV) of the 42 clonal values (i.e. estimated from the ramet
165 phenotypes) and of the 42 ortet phenotypic values adjusted for effects related to the experimental
166 design (see below) were computed for each trait as: $CV = \frac{\sigma}{\mu} \times 100$, with σ the standard deviation and
167 μ the mean value.

168

169 2.3. *Genotyping*

170 Molecular data were obtained by GBS [25,26] for the 42 ortets, 93 Deli and 91 La Mé parents
171 of the training hybrid crosses (Table 1). Ortets genotypes were obtained from two or three samples
172 collected on different ramets (thus allowing controlling the legitimacy of the ramets). DNA extraction
173 and GBS were performed as described in Cros et al. [15], using the *Pst*I and *Hha*I restriction enzymes.
174 The raw fastq sequence data were processed with Tassel GBS v. 5.2.44 [27], using the Bowtie2
175 software for alignment [28], and VCFtools 0.1.14 [29]. The indels were discarded, the datapoints with
176 depth below five were set to missing, the SNPs that were not biallelic, with more than 75% of missing
177 data or on the unassembled part of the genome were discarded (see Cros et al. [15] for more details
178 about SNP calling and filtering). This resulted in a dense genome covering with 15,054 SNPs. The
179 average percentage of missing data was 23.08% (3.64% - 43.42% per individual). To explain the
180 differences in accuracy between ASGM and PSAM, the distribution of the minor allele frequency

(MAF) and of the frequency of the alternate allele (i.e. that was not present on the reference genome) were computed in Deli and La Mé, as well as the correlation among populations for each of these two parameters.

2.4. *Imputation of missing SNP data and phasing*

Imputation of missing SNP data and phasing were carried out with Beagle 4.0 [30]. This software can consider the family relationships (i.e. parent-offspring) and infers missing genotypes using genotype likelihood computed from the pedigree. The process followed to impute and phase the SNP data is given in Fig. 1. The pedigree of the population involved in this study is available over several generations. For imputation, the initial SNP dataset containing all the genotyped individuals was divided into three distinct SNP datasets containing the Deli parents, the La Mé parents and the ortets, respectively. The Deli and La Mé SNP datasets were imputed separately giving to the software their respective pedigrees, and were then merged with the unimputed SNP dataset of ortets. The resulting global dataset was imputed and phased, providing the software with the pedigree file indicating the Deli and La Mé parent of each ortet. Nine ortets had one parent for which the DNA was unavailable but, as the missing parents were obtained through selfing, the selfed grandparent was used in the pedigree instead of the actual parent. For the other steps of the analysis that required a pedigree, the real pedigree was used.

2.5. *Definition of SNP datasets*

To quantify how the characteristics of the SNP dataset (i.e. maximum percentage of missing data allowed per SNP, p_{max}, and resulting number of SNPs, n_{snp}) affected the GS accuracy, we made genomic predictions using different SNP datasets with varying maximum percentage of missing data per SNP, as shown in table 2. Thereby, for the rest of the study, the SNP dataset will refer to an SNP matrix with a given number of SNPs resulting from the filtering made on the maximum percentage of missing data allowed per SNP.

207

208 2.6. *Prediction models and computation of genetic values of unobserved clones*

209 Two approaches were implemented to predict the genetic value of the validation clones: the
210 across-population SNP genotype model (ASGM) and the population-specific effects of SNP alleles
211 model (PSAM). In addition, for both approaches, two models were tested: a purely additive model
212 (ASGM_A and PSAM_A) and a model combining additive and dominance effects (ASGM_AD and
213 PSAM_AD). The ASGM_A approach used a model with a single random genetic effect,
214 corresponding to the additive genetic value of the parents of the training hybrid crosses and of the
215 validation clones. The ASGM_AD and PSAM_AD model also included a random dominance effect of
216 crosses and ortets. The PSAM_A approach used two random effects partitioning the additive genetic
217 values of each individual into two parts originating from Deli and La Mé alleles. All these four models
218 were implemented separately on each trait (univariate models). For GS, the GBLUP statistical
219 approach was used [31,32], and the corresponding models were termed G_ASGM_A, G_ASGM_AD,
220 G_PSAM_A, and G_PSAM_AD. In addition, to evaluate the usefulness of the SNP data, these four
221 models were implemented with pedigree data instead of SNPs (control PBLUP models, termed
222 P_ASGM_A, P_ASGM_AD, P_PSAM_A, and P_PSAM_AD).

223 In all cases, the models were trained with the phenotypic data of ALT hybrids and the genomic
224 data of their parents, and the genetic values of the 42 validation clones were predicted. For all the
225 models mentioned above, no phenotypic data of the validation clones were provided to the prediction
226 models. This corresponds to a breeding situation where predictions are made for immature individuals
227 (e.g. nursery plantlets belonging to crosses that were not evaluated in progeny-tests but were produced
228 by mating the best parents selected at the end of the progeny-tests). However, ortet selection can also
229 be made within the crosses evaluated in progeny tests. In this case, the ortet candidates have
230 phenotypic data records, which should be taken into consideration along with their SNP data when
231 predicting their clonal value. This was evaluated with the G_ASGM_A model, simply including the
232 adjusted phenotypic value of the validation ortets (see below) to the phenotypic dataset used to train
233 the model, and is referred to as the G_ASGM_A+pheno approach.

All GS analyses were run on a server of the CIRAD-UMR AGAP HPC data center of the South Green bioinformatics platform (http://www.southgreen.fr/), using a homemade R script.

2.6.1. *Across-population SNP genotype models (ASGM)*

The model used for the G_ASGM_AD approach was as follows:

$$y = \boldsymbol{X}\beta + \boldsymbol{Z_1}g_i + \boldsymbol{Z_2}g_{Deli\times LM} + \boldsymbol{Z_3}b + \boldsymbol{Z_4}p + \varepsilon$$

with: y the observed phenotypes of the training hybrid individuals, β the vector of fixed effects (phenotypic mean, trial effects, block effects and, for bunch production traits, age), g_i ~ N(0, $\boldsymbol{H_i}\sigma^2_{a_i}$) the individual additive genetic effects, $g_{Deli\times LM}$ ~ N(0, $\boldsymbol{H}_{Deli\times LM}\sigma^2_{d_{Deli\times LM}}$) the genetic dominance effects, b ~ N(0, $\boldsymbol{I}\sigma^2_b$) the incomplete block effect, and p ~ N(0, $\boldsymbol{I}\sigma^2_p$) the elementary plot effects. $\boldsymbol{X}$, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$, $\boldsymbol{Z_3}$ and $\boldsymbol{Z_4}$ are the incidence matrices associated to β, g_i, $g_{Deli\times LM}$, b and p respectively. $\boldsymbol{H_i}\sigma^2_{a_i}$ and $\boldsymbol{H}_{Deli\times LM}\sigma^2_{d_{Deli\times LM}}$ are the variance-covariance matrices associated with g_i and $g_{Deli\times LM}$, respectively. $\sigma^2_{a_i}$ and $\sigma^2_{d_{Deli\times LM}}$ are the additive and dominance variances, respectively. ε ~ N(0, $\boldsymbol{I}\sigma^2_\varepsilon$) is the vector of residual effects and $\boldsymbol{I}$ the identity matrix. To implement this model in practice, two specificities of our dataset had to be taken into account. First, a few parents of the training crosses were not genotyped (Table 1), and the $\boldsymbol{H_.}$ matrices had therefore to be made with the genealogical data of hybrid crosses with ungenotyped parents and with the SNP data of hybrid crosses with genotyped parents (computed with the SNP data of their parents, see below) and of the ortets. All $\boldsymbol{H_.}$ matrices subsequently in this paper will refer to matrices combining genealogical and genomic information. $\boldsymbol{H_i^{-1}}$ is the inverse of $\boldsymbol{H_i}$, computed according to Misztal et al. [33] as: $\boldsymbol{H_i^{-1}} = \boldsymbol{A_i^{-1}} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G_i^{-1}} - \boldsymbol{A_{i_{22}}^{-1}} \end{bmatrix}$, where $\boldsymbol{G_i^{-1}}$ and $\boldsymbol{A_{i_{22}}^{-1}}$ are the inverse of the realized and the genealogical additive relationship matrices, respectively, of the 42 ortets and the hybrid crosses with genotyped parents, and $\boldsymbol{A_i^{-1}}$ is the inverse of the genealogical relationship matrix of all hybrid crosses (i.e. the few with ungenotyped parents and the ones with genotyped parents) and the 42 ortets. Second, the phenotyped individuals constituting the hybrid crosses were not genotyped while they had to be

connected to the validation ortets through their genomic relationships (only the parents of the hybrids were genotyped, except a few parents that were not genotyped and for which the genealogical relationships were used, as explained above). To get genotypes for the hybrid crosses with genotyped parents, we computed for each cross the mean genotypes expected from the parental genotypes (i.e. for SNP j in cross i, the mean number of copies of the minor allele of SNP j expected to be found in the hybrid individuals of i), assuming this was relevant considering the relatively large number of individuals per cross (Table 1). The genomic additive relationship matrix $\boldsymbol{G}$ was obtained as: $\boldsymbol{G} = \frac{\boldsymbol{X}'\boldsymbol{X}}{2\sum_{l=1}^{n_{SNP}} p_l(1-p_l)}$, with $\boldsymbol{X} = \boldsymbol{Z} - \boldsymbol{P}$, $\boldsymbol{X}$' the transpose of matrix $\boldsymbol{X}$, $\boldsymbol{Z}$ the SNP matrix containing the number of copies of the minor allele at an SNP (ranging from 0 to 2), $\boldsymbol{P}$ a matrix given by $\boldsymbol{P} = 2(p_l - 0.5)$, and p_l the frequency of the minor allele at SNP l [34]. $\boldsymbol{H}_{Deli \times LM}$ is the dominance relationship matrix combining genomic dominance relationships between crosses with parents and clones, and genealogical dominance relationships between the few crosses with ungenotyped parents. $\boldsymbol{H}^{-1}_{\boldsymbol{Deli \times LM}}$ was computed following the same method as $\boldsymbol{H}_{\boldsymbol{i}}^{-1}$ except that the additive relationship matrices were replaced by the dominance relationship matrices. The realized dominance relationship matrix $\boldsymbol{G_D}$ was computed according to Su et al. [35] as: $\boldsymbol{G_D} = \frac{\boldsymbol{\Pi}\boldsymbol{\Pi}'}{2\sum p_l q_l (1-2p_l q_l)}$, with $\boldsymbol{\Pi}$ the $n \times m$ matrix (n: number of hybrid crosses and clones and m: number of SNPs) of heterozygosity coefficients with element $\boldsymbol{\Pi}_{kl} = 0 - p_l q_l$ if clone or ortet k is homozygous and $\boldsymbol{\Pi}_{kl} = 1 - p_l q_l$ if it is heterozygous at locus l, and p_l and q_l the frequencies of the first and the second allele at locus l. The purely additive approach ASGM_A used the same model without the dominance effect.

For the P_ASGM_A and P_ASGM_AD, $\boldsymbol{H_i}$ was replaced by the additive genealogical relationship matrix $\boldsymbol{A_i}$ and, for P_ASGM_AD, $\boldsymbol{H_{Deli \times LM}}$ was replaced by the genealogical dominance relationship matrix.

The estimated genetic value for the validation clones was $\hat{g}_i$ and, for G_ASGM_AD and P_ASGM_AD, $\hat{g}_i + \hat{g}_{Deli \times LM}$.

2.6.2. *Population-specific effects of SNP alleles models (PSAM)*

The model used for G_PSAM_AD was as follows:

$$y = \boldsymbol{X}\beta + \boldsymbol{Z_1}g_{Deli} + \boldsymbol{Z_2}g_{LM} + \boldsymbol{Z_3}g_{Deli\times LM} + \boldsymbol{Z_4}b + \boldsymbol{Z_5}p + \varepsilon$$

with g_{Deli} ~ N(0, $\boldsymbol{H_{Deli}}\sigma^2_{g_{Deli}}$) and g_{LM} ~ N(0, $\boldsymbol{H_{LM}}\sigma^2_{g_{LM}}$) the additive effects inherited by the parents of the hybrid crosses and the ortets from the Deli and La Mé populations, respectively, and $g_{Deli\times LM}$ ~ N(0, $\boldsymbol{H}_{Deli\times LM}\sigma^2_{d_{Deli\times LM}}$) the dominance effects of the crosses and clones. $\boldsymbol{X}$, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$, $\boldsymbol{Z_3}$, $\boldsymbol{Z_4}$, $\boldsymbol{Z_5}$ are the incidence matrices associated to β, g_{Deli}, g_{LM}, $g_{Deli\times LM}$, b and p, respectively. $\boldsymbol{H_{Deli}}\sigma^2_{g_{Deli}}$, $\boldsymbol{H_{LM}}\sigma^2_{g_{LM}}$ and $\boldsymbol{H_{Deli\times LM}}\sigma^2_{d_{Deli\times LM}}$ are the variance-covariance matrices associated to g_{Deli}, g_{LM} and $g_{Deli\times LM}$, respectively. $\sigma^2_{g_{Deli}}$ and $\sigma^2_{g_{LM}}$ are the additive genetic variances of the Deli and La Mé populations, respectively, and $\sigma^2_{d_{Deli\times LM}}$ is the genetic dominance variance of crosses and clones. $\boldsymbol{H_{Deli}}$ is the matrix combining the additive realized relationships of the clones and the genotyped Deli parents of the crosses and the additive genealogical relationships of the few ungenotyped Deli parents of the hybrid crosses. $\boldsymbol{H_{LM}}$ is defined similarly for the La Mé population. $\boldsymbol{H_{Deli}}$ and $\boldsymbol{H_{LM}}$ were created following the same procedure as $\boldsymbol{H_i}$. For each parental population, the required realized relationship was computed according to VanRaden [34] (see above) except that in the SNP matrices ($\boldsymbol{Z_{Deli}}$ and $\boldsymbol{Z_{LM}}$) containing the number of copies of minor allele inherited from the considered parental population, the genotypes of clones were coded into 0 and 1, as indicated by the phase information provided by Beagle 4.0, while the genotypes of the hybrid's parents were coded into 0, 1, and 2, as in $\boldsymbol{Z}$. $\boldsymbol{H_{Deli\times LM}}$ is the dominance relationship matrix containing both realized dominance relationships between clones and crosses implying genotyped parents, and genealogical dominance relationships between the crosses implying ungenotyped parents, computed as: $\boldsymbol{H_{Deli\times LM}} = \boldsymbol{H_{Deli}} \otimes \boldsymbol{H_{LM}}$, with $\otimes$ the Kronecker product.

For P_PSAM_A and P_PSAM_AD, $\boldsymbol{H_{Deli}}$ and $\boldsymbol{H_{LM}}$ were replaced by the additive genealogical relationship matrices $\boldsymbol{A_{Deli}}$ and $\boldsymbol{A_{LM}}$ and, for P_PSAM_AD, $\boldsymbol{H_{Deli\times LM}}$ was replaced by the genealogical dominance relationship matrix.

309 The estimated genetic value for the validation clones was calculated as the sum of the additive
310 genetic values inherited from the two parents, i.e. $\hat{g}_{Deli} + \hat{g}_{LM}$ and, for G_PSAM_AD and
311 P_PSAM_AD, of its dominance value, i.e. $\hat{g}_{Deli} + \hat{g}_{LM} + \hat{g}_{Deli \times LM}$.

312

313 2.7. *Prediction accuracies*

314 The ability of each model to predict the reference clonal value of the 42 validation clones (see
315 below) was evaluated through their prediction accuracy, computed as the correlation between the
316 reference value and the predicted clonal values.

317 Pairwise comparisons of prediction accuracies among models were made for each trait using
318 the Hotelling–Williams t-test [36]. This test compares two non-independent correlations, i.e. having
319 one variable in common, which in our case is the reference value of the 42 clones. This test was
320 applied using the R package *psych* [37].

321

322 2.8. *Determination of the reference clonal values predicted by the models*

323 In order to validate the different prediction models, clonal genetic values were obtained for
324 each clone from the phenotypic data collected on their ramets. Subsequently in this paper, they will be
325 referred to as reference genetic values. They were computed using a simple linear mixed model to
326 adjust the phenotypic values of the ramets for the effects of experimental design, i.e. clonal trials,
327 blocks, incomplete blocks, elementary plots and, for bunch production traits, age. In this model, clones
328 were included as a fixed effect.

329

330 2.9. *Accuracy of phenotypic selection before clonal trials*

331 To evaluate the possibility of using GS instead of the current phenotypic selection (PS) to
332 select the hybrid individuals to test in the clonal trials, the PS accuracy was computed for each trait. It
333 was defined as the correlation between the ortet adjusted phenotypes and the reference clonal genetic

values. The adjusted phenotype was obtained for each ortet from its phenotypic data collected in AK1, using a simple linear mixed model with individuals as random effect and hybrid crosses and all the effects related to the experimental design, i.e. trials, blocks, incomplete blocks, elementary plots and, for bunch production traits, age, as fixed effects. Finally, each ortet had for each trait an adjusted phenotype that was equal to the sum of the individual effect of the ortet, the effect of its cross and the mean residual effect over its phenotypic data records.

3. Results

3.1. *Distribution of frequencies of minor and alternate alleles across population*

The distribution of MAF in both Deli and La Mé populations showed a reduction in the number of SNPs with the increase of MAF (Fig. 2). The MAF ranged from 0 to 0.5 for both La Mé and Deli populations and the average was 0.1 for La Mé (Fig. 2a) and 0.07 for Deli (Fig. 2b). Most SNPs had low MAF values (<0.05) in both populations. La Mé populations had 65.6% SNPs with MAF<0.05, against 73.3% SNPs in Deli (i.e. 11.7% more SNPs with low MAF in Deli). In contrast, fewer SNPs had high MAF (>0.40) in both populations, and they were higher in proportion in La Mé (8.2% SNPs) than in Deli (4.8%). This showed the lower genetic diversity of Deli parents compared to La Mé, which resulted from their contrasted history with more generations of selection, drift and inbreeding in Deli than in La Mé.

Correlation between La Mé and Deli MAF (Fig. 2c) shows SNPs largely concentrated alongside *x* and *y* axes, demonstrating that most SNPs have distinct segregation patterns among Deli and La Mé, i.e. being fixed or almost fixed in one population while segregating, and in many cases with a high MAF, in the other population. Thus, 31.5% of the SNPs were fixed or almost fixed in one population (MAF<0.05) while segregating with MAF≥0.05 in the other population. This is the result of the high genetic difference between Deli and La Mé populations, for which the *Fst* fixation index reaches 0.55 [38]. In detail, for these SNPs, MAF<0.05 was more often observed in Deli (19.6% of all SNPs had MAF<0.05 in Deli and MAF>=0.05 in La Mé) than in La Mé (11.9% of all SNPs had MAF<0.05 in La Mé and MAF>=0.05 in Deli), again as a result of the lower genetic diversity of the

Deli population. Also, the number of SNPs segregating with MAF>0.05 in both populations was low (14.8% of all SNPs). Despite these differences, a large number of SNPs (53.7% of all SNPs) had MAF<0.05 in both populations, showing segregation with rare alleles in both Deli and La Mé. However, correlation of the frequency of the alternate allele between La Mé and Deli (Fig. 2d) over all SNPs showed that 62.8% of SNPs have a frequency of alternate allele smaller than 0.05 in one population and greater than 0.95 in the other population, i.e. fixed or almost fixed in the two populations but for different alleles. Hence, given that most of the SNPs (85.2%) have either MAF<0.05 in one population and MAF>=0.05 in the other population (31.5%), or MAF<0.05 in both populations but for different alleles (53.7%), the use of PSAM is justified.

3.2. *Effect of GS prediction model and SNP dataset on prediction accuracy*

Prediction accuracies of GS methods ranged from -0.03 to 0.70 depending on prediction model, trait and SNP dataset (Fig. 3) for additive models (G_ASGM_A and G_PSAM_A). Indeed, in a preliminary analysis, inconsistent differences or similar accuracies were observed between additive models and additive + dominance models, depending on marker dataset and trait (see Supplementary Fig. S. 1). Henceforward, we will only refer to additive models.

On average over traits and SNP datasets, G_ASGM_A was more accurate (0.45) than G_PSAM_A (0.37), with the mean prediction accuracy per trait over SNP datasets ranging from 0.14 (PF) to 0.65 (FB) for G_ASGM_A and from 0.09 (PF) to 0.58 (FB) for G_PSAM_A. G_ASGM_A obtained a mean prediction accuracy greater than G_PSAM_A for six traits out of eight, with G_PSAM_A being on average slightly more accurate than G_ASGM_A for ABW and equal for BN (Table 3). Considering the maximum accuracy over all SNP datasets, the prediction accuracy ranged from 0.18 (PF) to 0.70 (FB) for G_ASGM_A and from 0.16 (PF) to 0.65 (FB) for G_PSAM_A (Table 3), and, here, G_PSAM_A was more accurate for both BN and ABW, although slightly. Considering the different SNP datasets and traits, large differences in prediction accuracy between G_ASGM_A and G_PSAM_A were observed, up to +0.36 in favour of G_ASGM_A with OP at p_{max} = 45%-n_{SNP} = 11,707 (Fig. 3 and Table 4). The differences in prediction accuracies between G_ASGM_A and

388 G_PSAM_A were significant for three traits in four cases (Table 4). Prediction accuracies of
389 G_ASGM_A were significantly greater than G_PSAM_A for OP with two SNP datasets (p_{max}=45%-
390 n_{SNP}=11,707 and p_{max}=75%-n_{SNP}=15,054), FB and FFB in one dataset each, p_{max}=10%-n_{SNP}=6,898 and
391 p_{max}=5%-n_{SNP}=5,620 respectively. In rare cases, low and non-significant differences (up to +0.16) were
392 observed in favor of G_PSAM_A. G_ASGM_A, therefore, appeared to be a better approach (i.e. more
393 accurate and easier to implement) for predicting clonal values for oil palm yield components.

394 Prediction accuracies were broadly improved for three traits (FB, BN and ABW) when
395 relationship matrices were computed using SNPs (G_ASGM_A and G_PSAM_A) instead of
396 genealogical data (control pedigree-based models P_ASGM_A and P_PSAM_A). The maximum
397 prediction accuracies of GS over all SNP datasets outperformed pedigree-based models for seven traits
398 out of eight (except for AFW) (Table 5). The largest difference was observed in BN for p_{max}=75%-
399 n_{snp}=15,054, with G_ASGM_A accuracy being 0.67 higher than P_ASGM_A. Accuracies of pedigree-
400 based models exceeded GS in almost every SNP dataset for AFW (Fig. 3 and Table 5). The
401 differences between GS models and their pedigree-based control models were significant for five
402 traits, with four traits (FB, OP, BN and ABW) where GS was the best and one trait (AFW) where
403 pedigree-based models were more accurate (Table 5).

404 The SNP dataset affected the prediction accuracy differently according to the trait and the
405 model. However, the use of a different SNP dataset for each combination of trait and model seems
406 unrealistic for the practical application of GS. Therefore, in order to identify the optimal SNP
407 dataset(s) that would maximize GS accuracy, we computed for each SNP dataset the mean
408 G_ASGM_A prediction accuracy over the traits. This value increased with the SNP density (0.41 with
409 SNP dataset p_{max}=0%-n_{snp}=2,447 and 0.43 with p_{max}=5%-n_{snp}=5,620), before plateauing at 0.46 with
410 the subsequent SNP datasets. Mean prediction accuracy over the SNP datasets forming the plateau
411 ranged from 0.17 (PF) to 0.66 (FB), and were close to the highest accuracies achieved over all the
412 SNP datasets (Table 3). There was therefore a minimum of 6,898 SNPs required to reach maximum
413 prediction accuracy on average over all traits.

414 Accuracies were more variable among SNP datasets and traits with G_PSAM_A than with
415 G_ASGM_A. With G_ASGM_A, prediction accuracies tended to increase with SNP density before

plateauing (except for AFW) and slightly decreasing in some cases. This suggested that more useful information was captured for prediction purposes when using more SNPs (to a certain limit) and, again, that the percentage of missing data was of lesser importance. On the other hand, a reduction of accuracies was observed with SNP density for AFW. For G_PSAM_A, prediction accuracies increased, and usually plateaued, for only four traits (FB, PF, NF and ABW). For the other traits, prediction accuracies remained stable or tended to decrease with increasing marker density and maximum percentage of missing SNP data. Thus, the accuracy of OP, for instance, decreased around 59.6% from p_{max}=0%-n_{snp}=1,497 to p_{max}=45%-n_{snp}=11,425 (Fig. 3).

3.3. *Comparison of prediction accuracies of PS and GS*

Figure 4 presents the prediction accuracies of PS and the mean prediction accuracy of G_ASGM_A over the best datasets (i.e. with p_{max} from 10% to 75% and n_{snp} from 6,898 to 15,054), with (G_ASGM_A+pheno) and without phenotypic data of the ortets. Variation of PS accuracy was large between traits, going from -0.03 for ABW to 0.63 for OP. Very low PS accuracies (<0.1) were obtained for ABW and FFB, meaning that PS would have been inefficient for these two traits. The highest PS accuracies were achieved in OP (0.63) and PF (0.59) (Table 6 and Fig. 4). These two traits are known to have moderate to high heritability in the oil palm [2] and are consequently routinely used for preselection before clonal trials. This was the case here, as indicated by the intensity of PS for these two traits, which was the highest among the eight traits studied (Table 6).

The GS prediction accuracy obtained with the best SNP datasets was generally higher with G_ASGM_A+pheno than with G_ASGM_A (except for AFW, where a slight decrease was found) (Fig. 4). On average over all the traits, G_ASGM_A+pheno thus reached 0.53, against 0.46 for G_ASGM_A (i.e. +15.2%). The prediction accuracy of G_ASGM_A and G_ASGM_A+pheno obtained with the best SNP datasets was above PS prediction accuracies for six and seven traits, respectively, out of eight. On average over all traits, the prediction accuracies of G_ASGM_A and G_ASGM_A+pheno were, respectively, 64.3% and 89.3% greater than PS (0.28). The case where GS outperformed PS the most was ABW with the G_ASGM_A+pheno model, with an accuracy of 0.62

against -0.03. PS only surpassed G_ASGM_A for two traits (PF and OP) and G_ASM_A+pheno for one trait (PF).

4. Discussion

In this paper, we evaluated the possibility of predicting the genetic value of oil palm ortet selection candidates, using GS models and high throughput SNP genotyping (GBS). We considered two breeding situations consisting of candidate ortets with or without phenotypic values. We assessed the effect on prediction accuracy of marker datasets and of two approaches for modeling the parental origin of marker alleles (across-population SNP genotype models, ASGM, and population-specific effects of SNP alleles models, PSAM).

4.1. *Improving the genetic progress of clonal breeding with GS*

In the current clonal breeding methodology, ortets that will be evaluated in clonal trials are selected on the few traits with high H^2 value among a limited number of phenotyped candidates at the mature stage and belonging to the best crosses evaluated in progeny tests. Based on the results presented here, annual genetic progress can be improved by selecting ortets (1) among a large population of the best possible crosses (produced based on the results of the progeny tests) at the juvenile (e.g. nursery) stage with GS models on most of the yield components or, (2) at the mature stage on all the yield components, using jointly the genomic and phenotypic data of the ortet selection candidates.

In detail, in the first GS approach that is now possible, the best crosses identified based on the results of the progeny test (i.e. with the best performance expected from the parental GCAs and the crosses' specific combining abilities [SCAs]) would be produced to generate a large number of seedlings, that would be submitted to GS on the traits with satisfactory GS accuracy. This would improve the genetic progress at three levels. First, most of the breeding programs consider that there are six traits of interest for palm oil yield breeding (FB, PF, OP, ABW, BN and FFB), and PS before

clonal trials is usually applied to PF and OP, as they have the highest H^2 [39]. In our dataset, these traits indeed had high H^2, with PS prediction accuracy >0.5 (Fig. 4) (although it was not clear why FB had a similar H^2, while it is usually among the traits with low H^2). Therefore, considering that breeders use 0.5 as the minimum prediction accuracy for applying PS before clonal trials, they would now apply GS to four traits (FB, OP, FFB and ABW) (Fig. 4), with a similar mean prediction accuracy over these traits with GS (0.56) compared to PS (0.60 over FB, PF and OP). Interestingly, the two traits that had a prediction accuracy lower with G_ASGM_A than with PS, i.e. PF and OP, were the ones for which the 42 ortets were submitted to the strongest phenotypic selection before clonal trials. In particular, PF had the highest intensity of phenotypic selection (0.68) and also had much lower prediction accuracy with G_ASGM_A than with PS. We hypothesized this occurred as the phenotypic preselection led to the fixation of many genes controlling these traits, and in particular PF, in the 42 ortets, thus making that the relationships computed over the genome-wide SNPs no longer matched with the relationships at the genes. This hypothesis could be investigated using a validation set that was not submitted to phenotypic preselection. Such a study would be of great interest as, in case our hypothesis could be confirmed, the breeders would likely get in practice a higher GS accuracy for PF and OP, as the seedlings comprising the population of application would not be preselected. In this case, GS before the clonal trials would be even more useful. Second, a GS-based approach would also increase the genetic progress by higher selection intensity compared to PS: GS would be applied to nursery individuals, i.e. possibly in the thousands, while PS is currently applied to the small number of individuals planted in the progeny tests trials (i.e. normally 10 to 50 per cross) [9]. Third, making the selection in the best possible crosses instead of the best crosses evaluated would be an improvement in terms of genetic progress, as the best possible crosses were likely not present in the progeny tests, due to the high degree of incompleteness of the mating designs. It is also possible to make these crosses in the context of phenotypic clonal selection, but in this case, the selection process would require around 10 more years of phenotypic evaluations in these elite crosses to identify the candidate ortets for the clonal trials [16].

In the second GS approach, i.e. the selection of ortets among mature hybrid individuals, it is now possible to apply this selection to all the yield components. Indeed, for individuals at the mature

stage, which thus may have phenotypic records, for each of the six commonly selected oil yield components it is possible to reach a prediction accuracy of 0.5 (or almost, in the case of BN), using conventional PS for PF and G_ASGM_A+pheno for the other traits. In practice, increasing the number of traits on which ortets are selected before clonal trials will increase selection intensity and thus the genetic progress.

Another possible approach to improve the genetic progress would be to use genomic predictions to identify, before the progeny tests, the best possible crosses, and to use them to implement the first approach of clonal GS suggested here. For that purpose, progeny tests from the previous cycle could be used as a training population, and genomic ortet selection would be applied at the nursery stage in the best possible crosses. This approach would, therefore, have the additional advantage of shortening the breeding cycle (as it makes it possible to run the clonal trials simultaneously with the progeny tests), but it should be investigated in greater details as its efficiency also depends on the accuracy of the genomic estimated breeding values of the parents.

4.2. *Effects of prediction model and SNP dataset on prediction accuracies*

G_PSAM_A can model genetic differences between Deli and La Mé populations, as it considers population-specific SNP variances and SNP effects. For that reason, we expected G_PSAM_A to perform better than G_ASGM_A for many traits, considering the marked genetic difference between Deli and La Mé, with F_{st} around 0.55 [38]. However, G_PSAM_A only performed better than G_ASGM_A for BN and, to a lesser extent, ABW. We hypothesized that this was the consequence of stronger differences among Deli and La Mé populations in terms of QTLs for BN and ABW than of QTLs controlling the other traits. This makes sense when considering that Deli and La Mé belong to different heterotic groups defined based on their phenotypic values for BN and ABW, in which they have opposite and complementary characteristics. This is in agreement with the results of Tisné et al. [40], who found a large majority of distinct significant QTLs among groups A and B on bunch production traits, i.e. six in group A and ten in group B, against only one common QTL. This is also in agreement with the fact that a large part of the SNPs in the two populations have opposite minor alleles, with differences as extreme as having one allele fixed in one population and the other

allele fixed in the other population (Fig. 2b, c). However, not all SNPs showed these types of differences and similar segregation patterns among populations were also observed, which is likely related to the similar performance of G_ASGM_A and G_PSAM_A for the other traits. In order to help to understand the results obtained here, it would be useful to investigate whether the QTLs identified in other studies for the different traits are located in regions of the genome where SNPs have similar or contrasted segregation. Also, it would be interesting to compare, across the Deli and La Mé populations, the linkage phases between SNP markers and the SNP effects, as it was previously done in cattle and maize [41]

Although G_PSAM_A has the potential to model genetic differences between parental populations, it also has a drawback, which is that it has to estimate more parameters than G_ASGM_A (i.e. more genetic variances and, because additive effects are split into two parts inherited from the two parental populations, more genetic effects) [42]. For example, while for a given clone a single genetic effect is estimated with G_ASGM_A, two genetic effects, i.e. one for each of the hybrid parents, are estimated with P_ASGM_A. Our results corroborate those of Zeng et al. [42] who attributed low accuracies in many scenarios of PSAM in animal studies to the complexity of the model caused by the segregation of SNP in the two parental breeds, and the resulting need to estimate two substitution effects per SNP instead of one.

Ibánez-Escriche et al. [20] obtained a significant advantage of G_PSAM_A over G_ASGM_A on accuracy for a low marker density (400 markers), a large number of records in the training population (4,000) and a relationship between breeds that was weak (i.e. common origin 550 generations ago) or absent. Similarly, Esfandyari et al. [43] found that G_PSAM_A outperformed G_ASGM_A for genetically distant hybrid parents, i.e. having diverged 300 to 400 generations ago, and a large training population with 2,000 to 8,000 individuals. The small advantage of G_PSAM_A over G_ASGM_A obtained in our study might, therefore, result from the fact that the genetic difference between the Deli and La Mé populations was actually not large enough (the Deli also having African ancestors, planted in Indonesia in 1848) and/or because of our training population was too small. Technow et al. [22] found higher accuracy while using G_PSAM_A+D than when using G_ASGM_A+D, with the gain in accuracy being larger with low SNP density (from 0.3 to 1 SNP per

megabase pair, Mbp) than with high marker density (10 SNP per Mbp). Here, considering the length of the oil palm genome is 1.8 Gb [44], the investigated range of SNP density was similar, going from 0.8 to 8.4 SNP per Mbp. Moreover, Lopes et al. [45] obtained similar prediction accuracies between G_ASGM_A and G_PSAM_A with high SNP density (31,930 SNPs). We did not find SNP density to have such an effect on the prediction accuracy of G_PSAM_A or on the relative performance of G_PSAM_A and G_ASGM_A. This likely results from the fact that, in our study, SNP density varied with SNP quality, with higher SNP numbers meaning a higher percentage of missing data. These two parameters, therefore, seem to interact on the prediction accuracy of the two models investigated. However, the fact that the mean GS accuracy over the traits increased with the number of SNPs and plateaued from 6,898 SNPs indicated that SNP density was of greater importance for the prediction accuracy than the percentage of missing data per SNP.

We found that, in order to maximize the efficiency of GS, the prediction of the genetic values must be done using G_ASGM_A with an SNP density ranging from around 7,000 to 15,000 for all traits. Another possibility would be to use a different SNP dataset for each trait, maximizing the accuracy for the considered trait. However, as previously mentioned, this does not seem convenient for the practical application of GS. The variation in prediction accuracy among SNP datasets might also have been exacerbated by the small size of our validation population (due to the difficulty of obtaining a large number of clones in trials, mainly because of the mantled anomaly [8]), and therefore so far it seems wiser to identify the best SNP datasets on average over several traits.

GS prediction models (G_ASGM_A and G_PSAM_A) were usually more accurate than their respective control pedigree-based models (P_ASGM_A and P_PSAM_A). The superiority of GS models shows that, even for unobserved individuals, GS models can account for both Mendelian sampling terms of siblings in a family and for family effects, while pedigree-based models can only account, at best, for family effects, as already found in previous oil palm GS studies [16].

However, G_ASGM_A outperformed its control pedigree-based model more often than G_PSAM_A. Thus, G_PSAM_A remained less accurate than P_PSAM_A for all the SNP datasets in three traits, while that never happened with G_ASGM_A. Also, the overall inferiority of G_PSAM_A

579 to G_ASGM_A occurred while P_PSAM_A was actually better than P_ASGM_A for five traits out of
580 eight. This looks contradictory and suggests that the performance of G_PSAM_A could have been
581 reduced by phasing errors. Also, many studies comparing G_ASGM_A and G_PSAM_A were carried
582 out by simulation with known phases [22,42,43], and therefore possible phasing errors in our study
583 could also be the cause of the discrepancies observed between our results and the results obtained in
584 simulation studies. Investigating other phasing approaches seems therefore of interest in the oil palm
585 context.
586
587 4.3. *Genotyped individuals for training*
588 In this study, to make GS predictions more cost-effective, the genotypes of the phenotyped
589 hybrid individuals constituting the training set were reconstructed using the molecular data of their
590 parents, with G_ASGM, or not used in the model, with G_PSAM. Both modeling approaches
591 therefore assume that the mean genotype in a hybrid family (i.e. the mean number of copies of the
592 minor allele over the individuals making the family) expected from the parental genotypes is the same
593 as the actual mean genotype. Nevertheless, in the case of allele segregation distortion at a locus, the
594 mean genotype in a hybrid family would significantly deviate from the mean genotype expected from
595 the parental genotypes, and this could reduce the GS accuracy. Indeed, high numbers of distorted
596 markers can be found in plants: Zuo et al. [46] and Li et al. [47] found more than 10% of markers
597 (SNP and SSR) significantly distorted. For future studies, it would be of great interest to compare the
598 approach used here with predictions made using real hybrid genotypes, and to measure the differences
599 in terms of GS accuracy and cost.
600
601 4.4. *Prediction of dominance effects*
602 GS prediction accuracies were not significantly enhanced by adding dominance effects.
603 Including dominance effects in the statistical model sometimes slightly increased or reduced
604 accuracies, depending on the traits and the SNP datasets, revealing a negligible genetic dominance
605 variance captured by the model compared to the total genetic variance, as already observed with

606 genomic predictions for performances of oil palm hybrid crosses [15] We assume this was a
607 consequence of reciprocal recurrent selection, which generated the contrasted allele frequencies we
608 observed across Deli and La Mé populations (Fig. 2), thus decreasing the ratio of SCA variance to
609 GCA variance [48] and making dominance effects absorbed by the GCAs or the population mean [41]

610

611 ## 5. Conclusion

612 This work showed that GS can largely improve clonal selection in oil palm (*Elaeis*
613 *guineensis*). GS prediction accuracies for ortets without phenotypic data records extended from -0.03
614 to 0.7 according to the trait, GS model and SNP dataset. The G_ASGM_A approach was better for
615 predicting clonal values than G_PSAM_A (more robust over traits and SNP datasets, easier to
616 implement), although G_PSAM_A could, in some cases, slightly improve prediction accuracies for the
617 two traits defining the heterotic groups. G_ASGM_A gave higher prediction accuracies than current
618 phenotypic selection for six traits out of eight. GS models required at least 7,000 SNPs to perform
619 best, with the percentage of missing data per SNP being of secondary importance.

620 The annual genetic progress of clonal oil palm breeding for yield can be increased by
621 replacing the current phenotypic ortet preselection before clonal trials by (1) genomic ortet
622 preselection on most of the yield components among a large population of the best possible crosses
623 (produced based on the results of the progeny tests) at the juvenile stage or, (2) ortet preselection at the
624 mature stage on all the yield components using jointly the genomic and phenotypic data of the ortet
625 selection candidates. GS can, therefore, enhance oil palm production. Further studies should be
626 conducted, for example considering other traits (vegetative growth, resistance to diseases) and using a
627 different phasing approach.

628

629 ## Acknowledgments

630 The authors acknowledge SOCFINDO (Indonesia), CRAPP (Benin) and PalmElit (France) for
631 planning and carrying out the field trials with CIRAD (France) and authorizing the use of the

phenotypic data for this study. We thank Bertrand Pitollat (CIRAD) for help in cluster management and Nicolas Turnbull (PalmElit) for leaf sample collection in clonal trials. We acknowledge the CETIC (African Center of Excellence in Information and Communication Technologies) for its support, and we thank the UMR AGAP genotyping technology platform (CIRAD, Montpellier), the DArT company (www.diversityarrays.com) and the CIRAD-UMR AGAP HPC data center of the South Green bioinformatics platform (http://www.southgreen.fr/) for their help. This research was partly funded by a grant from PalmElit SAS.

Data availability

The datasets are available from the corresponding author on reasonable request and with the permission of PalmElit.

Conflict of interests

The authors declare no conflict of interest.

Author contributions

AN carried out data analysis, with the help of DC. The paper was written by AN and DC, with the help of FJ and JMB. IS, DA and LN provided assistance and logistics for producing the plant material, managing field trials and collecting phenotypic data. BC, TDG, IS and DA designed field experiments. The molecular data were generated by AM, VR and VP.

References

[1] USDA, http://www.fas.usda.gov/data/oilseeds-world-markets-and-trade. Accessed January 13, 2020., 2020.

[3] J.P. Gascon, C. Berchoux, 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 (1964) 75–84.

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

[5] A. Rival, P. Levang, Palms of controversies: oil palm and development challenges, CIFOR, Jakarta, Indonésie, 2014. http://www.cifor.org/publications/pdf_files/Books/BLevang1401.pdf (accessed October 23, 2014).

[6] R. Corley, I. Law, The future for oil palm clones, in: 1997: pp. 279–289.

[7] E. Jaligot, A. Rival, T. Beulé, S. Dussert, J.-L. Verdeil, Somaclonal variation in oil palm (Elaeis guineensis Jacq.): the DNA methylation hypothesis, Plant Cell Reports. 19 (2000) 684–690.

[8] M. Ong-Abdullah, J.M. Ordway, N. Jiang, S. Ooi, S.-Y. Kok, N. Sarpan, N. Azimi, A.T. Hashim, Z. Ishak, S.K. Rosli, Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm, Nature. 525 (2015) 533.

[9] A.C. Soh, S. Mayes, J.A. Roberts, Oil Palm Breeding: Genetics and Genomics, CRC Press, 2017.

[10] B. Nouy, J.-C. Jacquemard, E. Suryana, F. Potier, K. Konan, T. Durand-Gasselin, The expected and observed characteristics of several oil palm (# Elaeis guineensis# Jacq.) clones, (2006).

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

[12] M.E. Goddard, B.J. Hayes, Genomic selection, J. Anim. Breed. Genet. 124 (2007). https://doi.org/10.1111/j.1439-0388.2007.00702.x.

[13] D. Grattapaglia, O.B. Silva-Junior, R.T. Resende, E.P. Cappa, B.S. Müller, B. Tan, F. Isik, B. Ratcliffe, Y.A. El-Kassaby, Quantitative genetics and genomics converge to accelerate forest tree breeding, Frontiers in Plant Science. 9 (2018) 1693.

[14] C. Wong, R. Bernardo, Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations, Theoretical and Applied Genetics. 116 (2008) 815–824.

[15] D. Cros, S. Bocs, V. Riou, E. Ortega-Abboud, S. Tisné, X. Argout, V. Pomiès, L. Nodichao, Z. Lubis, B. Cochard, Genomic preselection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses, BMC Genomics. 18 (2017) 839. https://doi.org/10.1186/s12864-017-4179-3.

[16] A. Nyouma, J.M. Bell, F. Jacob, D. Cros, From mass selection to genomic selection: one century of breeding for quantitative yield components of oil palm (Elaeis guineensis Jacq.), Tree Genetics & Genomes. 15 (2019) 69. https://doi.org/10.1007/s11295-019-1373-2.

[17] Q.B. Kwong, A.L. Ong, C.K. Teh, F.T. Chew, M. Tammi, S. Mayes, H. Kulaveerasingam, S.H. Yeoh, J.A. Harikrishna, D.R. Appleton, Genomic selection in commercial perennial crops: applicability and improvement in oil palm (Elaeis Guineensis Jacq.), Scientific Reports. 7 (2017) 2872.

[18] R. Durán, F. Isik, J. Zapata-Valenzuela, C. Balocchi, S. Valenzuela, Genomic predictions of breeding values in a cloned Eucalyptus globulus population in Chile, Tree Genetics & Genomes. 13 (2017) 74.

[19] D. Cros, L. Mbo-Nkoulou, J.M. Bell, J. Oum, A. Masson, M. Soumahoro, D.M. Tran, Z. Achour, V. Le Guen, A. Clement-Demange, Within-family genomic selection in rubber tree (Hevea brasiliensis) increases genetic gain for rubber production, Industrial Crops and Products. 138 (2019) 111464. https://doi.org/10.1016/j.indcrop.2019.111464.

[20] N. Ibánez-Escriche, R. Fernando, A. Toosi, J. Dekkers, Genomic selection of purebreds for crossbred performance, Genetics Selection Evolution. 41 (2009) 12.

[21] C. Stuber, C.C. Cockerham, Gene effects and variances in hybrid populations, Genetics. 54 (1966) 1279.

[22] F. Technow, C. Riedelsheimer, TobiasA. Schrag, AlbrechtE. Melchinger, Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects, Theor Appl Genet. 125 (2012) 1181–1194. https://doi.org/10.1007/s00122-012-1905-8.

Corley, R.H.V., Tinker, P.B., 2016. The oil palm, 5th ed. Wiley-Blackwell, Chichester, UK. https://doi.org/10.1002/9781118953297

[24] F. Potier, B. Nouy, A. Flori, J. Jacquarmard, H. Edyana Suryna, T. Durand-Gasselin, Yield potential of oil palm (Elaeis guineensis Jacq) clones: preliminary results observed in the Aek Loba genetic block in Indonesia, in: 2006.

[25] J. He, X. Zhao, A. Laroche, Z.-X. Lu, H. Liu, Z. Li, Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding, Frontiers in Plant Science. 5 (2014) 484. https://doi.org/10.3389/fpls.2014.00484.

[26] R.J. Elshire, J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, S.E. Mitchell, A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species, PLoS One. 6 (2011) e19379.

[27] J.C. Glaubitz, T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, E.S. Buckler, TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline, PloS One. 9 (2014) e90346.

[28] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nature Methods. 9 (2012) 357.

[29] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, The variant call format and VCFtools, Bioinformatics. 27 (2011) 2156–2158.

[30] S.R. Browning, B.L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering, The American Journal of Human Genetics. 81 (2007) 1084–1097.

[31] S.A. Clark, J. van der Werf, Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values, in: Genome-Wide Association Studies and Genomic Prediction, Springer, 2013: pp. 321–330.

[32] D. Habier, R. Fernando, J. Dekkers, The impact of genetic relationship information on genome-assisted breeding values, Genetics. 177 (2007) 2389–2397.

[33] I. Misztal, I. Aguilar, D. Johnson, A. Legarra, S. Tsuruta, T. Lawlor, A unified approach to utilize phenotypic, full pedigree and genomic information for a genetic evaluation of Holstein final score, Interbull Bulletin. (2009) 240.

[34] P.M. VanRaden, Efficient methods to compute genomic predictions, Journal of Dairy Science. 91 (2008) 4414–4423.

[35] G. Su, O.F. Christensen, T. Ostersen, M. Henryon, M.S. Lund, Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers, PloS One. 7 (2012) e45293.

[36] J.H. Steiger, Tests for comparing elements of a correlation matrix., Psychological Bulletin. 87 (1980) 245.

[37] W. Revelle, psych: Procedures for Psychological, Psychometric, and Personality Research, Northwestern University, Evanston, Illinois, 2018. https://CRAN.R-project.org/package=psych.

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

[39] R.H.V. Corley, P.B. Tinker, Vegetative Propagation and Biotechnology, in: The Oil Palm, John Wiley & Sons, Ltd, 2016: pp. 208–224. https://doi.org/10.1002/9781118953297.ch7.

[40] S. Tisné, M. Denis, D. Cros, V. Pomiès, V. Riou, I. Syahputra, A. Omoré, T. Durand-Gasselin, J.-M. Bouvet, B. Cochard, Mixed model approach for IBD-based QTL mapping in a complex oil palm pedigree, BMC Genomics. 16 (2015) 798.

[41] F. Technow, T.A. Schrag, W. Schipprack, E. Bauer, H. Simianer, A.E. Melchinger, Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize, Genetics. 197 (2014) 1343. https://doi.org/10.1534/genetics.114.165860.

[42] J. Zeng, A. Toosi, R.L. Fernando, J.C. Dekkers, D.J. Garrick, Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action, Genetics Selection Evolution. 45 (2013) 11.
[43] H. Esfandyari, A.C. Sørensen, P. Bijma, A crossbred reference population can improve the response to genomic selection for crossbred performance, Genetics Selection Evolution. 47 (2015) 76.
[44] R. Singh, M. Ong-Abdullah, E.-T.L. Low, M.A.A. Manaf, R. Rosli, R. Nookiah, L.C.-L. Ooi, S. Ooi, K.-L. Chan, M.A. Halim, Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds, Nature. 500 (2013) 335.
[45] M.S. Lopes, H. Bovenhuis, A.M. Hidalgo, J.A. Van Arendonk, E.F. Knol, J.W. Bastiaansen, Genomic selection for crossbred performance accounting for breed-specific effects, Genetics Selection Evolution. 49 (2017) 51.
[46] J.-F. Zuo, Y. Niu, P. Cheng, J.-Y. Feng, S.-F. Han, Y.-H. Zhang, G. Shu, Y. Wang, Y.-M. Zhang, Effect of marker segregation distortion on high density linkage map construction and QTL mapping in Soybean (Glycine max L.), Heredity. (2019). https://doi.org/10.1038/s41437-019-0238-7.
[47] C. Li, G. Bai, S. Chao, Z. Wang, A high-density SNP and SSR consensus map reveals segregation distortion regions in wheat, BioMed Research International. 2015 (2015).
[48] J.C. Reif, F.-M. Gumpert, S. Fischer, A.E. Melchinger, Impact of interpopulation divergence on additive and dominance variance in hybrid populations, Genetics. 176 (2007) 1931–1934.

Tables

Table 1

Characteristics of the datasets used for training and validation.

	Hybrid crosses (training set)		Hybrid clones (validation set)	
	bunch production	bunch quality	bunch production	bunch quality
Number of crosses or ortets	295	279	42	42
Number of individuals or ramets	19,668	12,341	2,908	1,439
Average number of individuals per cross or ramets per clone (min–max)	67 (17-503)	44 (21-274)	69 (5-138)	34 (4-74)
Number of Deli parents (genotyped)	108 (93)	103 (90)	16	16
Number of La Mé parents (genotyped)	102 (91)	100 (89)	12	12
Age at time of data collection (years)	3-7	5-9	3-7	5-9

Table 2

Characteristics of the SNP datasets defined based on a threshold in terms of maximum percentage of missing data per individual.

	Maximum percentage of missing data allowed per SNP p_{max} (resulting average)					
	0 (0)	5 (1.03)	10 (2.19)	25 (5.92)	45 (12.10)	75 (23.08)
Average percentage of missing data per individual in La Mé	0	1.49	3.20	8.81	15.31	23.95
Average percentage of missing data per individual in Deli	0	0.87	1.83	4.76	10.62	22.56
Number of SNPs n_{snp}	2,447	5,620	6,898	9,205	11,707	15,054

787 **Table 3**

788 Mean prediction accuracies according to trait and prediction model.

789 Bunch production: bunch number (BN), average bunch weight (ABW) and total bunch production
790 (FFB); bunch quality: average fruit weight (AFW), fruit to bunch (FB), pulp to fruit (PF), and oil to
791 pulp (OP) ratios, and number of fruits per bunch (NF); genomic prediction models: across-population
792 SNP genotype models (ASGM_A), population-specific effects of SNP alleles models (PSAM_A).
793 Values in brackets indicate the corresponding SNP dataset, defined on its maximum percentage of
794 missing data

795

Traits	**Mean accuracies over all SNP datasets**		**Maximum accuracies over all SNP datasets**	
	G_ASGM_A	G_PSAM_A	G_ASGM_A	G_PSAM_A
AFW	0.48	0.41	0.57 (0%)	0.49 (10%)
FB	0.65	0.58	0.70 (25%)	0.65 (75%)
PF	0.14	0.09	0.18 (45%)	0.16 (10%/75%)
OP	0.52	0.35	0.55 (45%)	0.47 (0%)
NF	0.47	0.43	0.54 (75%)	0.49 (75%)
FFB	0.47	0.30	0.55 (10%)	0.31 (10%)
BN	0.31	0.31	0.37 (75%)	0.40 (0%)
ABW	0.53	0.54	0.58 (75%)	0.60 (25%)
Mean	0.45	0.37	0.51	0.45

798

Table 4

Pairwise comparison of prediction accuracies among genomic selection and pedigree-based models, according to SNP dataset and trait. For any pair of models, the values indicate the difference in prediction accuracy between the two models (*model1 – model2*). SNP datasets are defined based on the maximum percentage of missing data allowed per SNP p_{max} and the resulting number of SNPs n_{SNP} and are labeled p_{max}%-n_{SNP}. Significance of pairwise comparisons by Hotelling–Williams t-test: *0.05 > P ≥ 0.01; **0.01 > P ≥ 0.001; ***P < 0.001.

SNP dataset	Compared models	AFW	FB	PF	OP	NF	FFB	BN	ABW
	P_ASGM_A – P_PSAM_A	-0.06	0.15*	0.06	-0.03	-0.04	0.03	-0.25**	-0.04
0%-2,447	*G_ASGM_A – G_PSAM_A*	0.15	0.08	0.12	0.07	0.16	0.01	-0.16	0.06
5%-5,620	*G_ASGM_A – G_PSAM_A*	0.06	0.06	0.06	0.04	-0.02	0.24*	0.01	-0.02
10%-6,898	*G_ASGM_A - G PSAM_A*	0.02	0.12*	0.00	0.06	0.02	0.23	-0.02	-0.02
25%-9,205	*G_ASGM_A - G PSAM_A*	0.11	0.09	0.10	0.14	0.02	0.22	0.08	-0.05
45%-11,707	*G_ASGM_A – G_PSAM_A*	0.01	0.12	0.05	0.36**	0.01	0.19	0.08	-0.02
75%-15,054	*G_ASGM_A - G PSAM_A*	0.10	-0.05	0.01	0.33*	0.04	0.16	-0.02	0.00

Table 5

Pairwise comparison of prediction accuracies among genomic selection and pedigree-based models, according to SNP dataset and trait. For any pair of models, the values indicate the difference in prediction accuracy between the two models (*model1 – model2*). SNP datasets are defined based on the maximum percentage of missing data allowed per SNP p_{max} and the resulting number of SNPs n_{SNP} and are labeled p_{max}%-n_{SNP}. Significance of pairwise comparisons by Hotelling–Williams t-test: *0.05 > P ≥ 0.01; **0.01 > P ≥ 0.001; ***P < 0.001.

SNP dataset	Compared models	AFW	FB	PF	OP	NF	FFB	BN	ABW
0%-2,447	*P_ASGM_A – G_ASGM_A*	-0.04	-0.12	0.00	-0.17	-0.01	0.07	-0.53**	-0.19
	P_PSAM_A – G_PSAM_A	0.16	-0.18	0.06	-0.07	0.19	0.05	-0.45*	-0.08
5%-5,620	*P_ASGM_A – G_ASGM_A*	0.03	-0.14	-0.01	-0.09	-0.01	-0.18	-0.56**	-0.28*
	P_PSAM_A – G_PSAM_A	0.14	-0.22*	-0.01	-0.02	0.03	0.04	-0.30	-0.25
10%-6,898	*P_ASGM_A – G_ASGM_A*	0.02	-0.20*	-0.07	-0.13	-0.01	-0.18	-0.59**	-0.30*
	P_PSAM_A – G_PSAM_A	0.09	-0.23*	-0.13	-0.04	0.05	0.02	-0.36*	-0.28*
25%-9,059	*P_ASGM_A – G_ASGM_A*	0.08	-0.20*	-0.08	-0.15	-0.02	-0.16	-0.64***	-0.30**
	P_PSAM_A – G_PSAM_A	0.24	-0.26*	-0.04	0.03	0.04	0.04	-0.30*	-0.31*
45%-11,425	*P_ASGM_A – G_ASGM_A*	0.11	-0.15	-0.09	-0.18*	0.03	-0.13	-0.62***	-0.30**
	P_PSAM_A –	0.17	-0.19	-0.10	0.22	0.08	0.04	-0.29	-0.28*

	G_PSAM_A								
75%-15,054	*P_ASGM_A – G_ASGM_A*	0.10*	-0.11	-0.08	-0.17	-0.08	-0.09	-0.67***	-0.34***
	P_PSAM_A – G_PSAM_A	0.26	-0.31**	-0.13	0.19	0.01	0.05	-0.44*	-0.30*

812

Table 6

Intensity and accuracy of phenotypic selection before clonal trials according to trait.

Traits	Intensity of selection	Phenotypic prediction accuracies
AFW	0.11	0.18
FB	0.32	0.59
PF	0.68	0.59
OP	0.58	0.63
NF	-0.27	0.46
FFB	0.19	0.09
BN	0.23	0.25
ABW	-0.01	-0.03

Figures

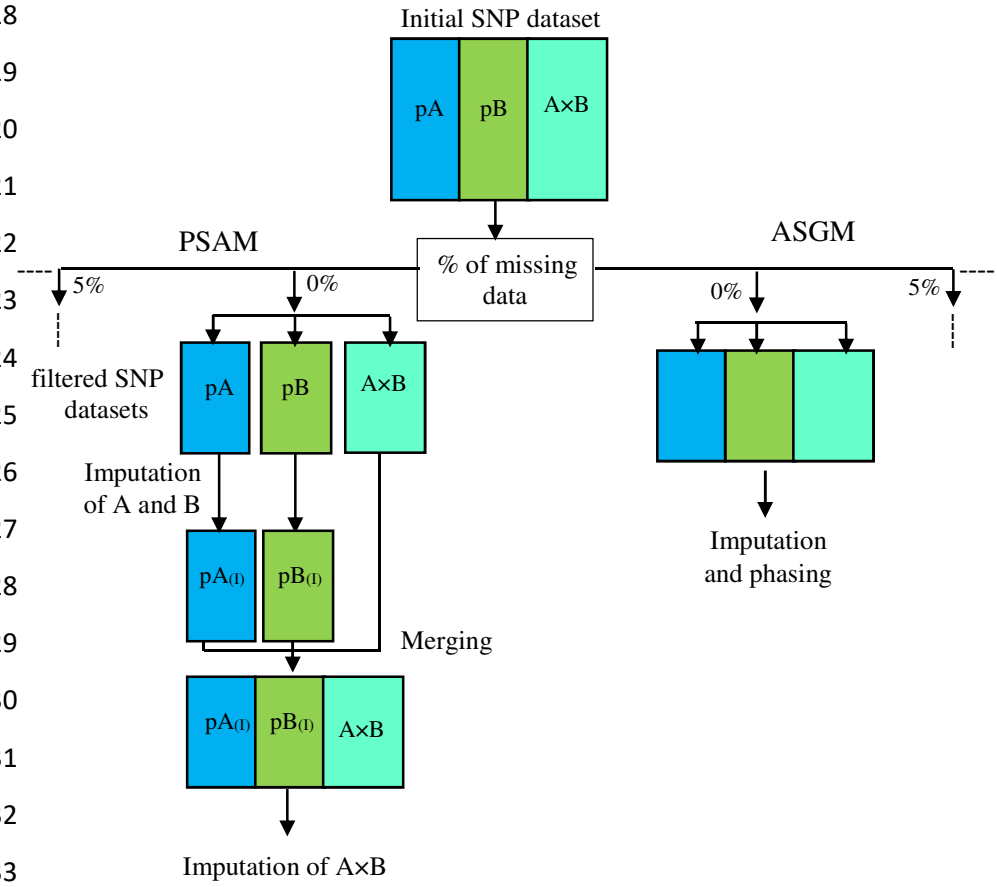
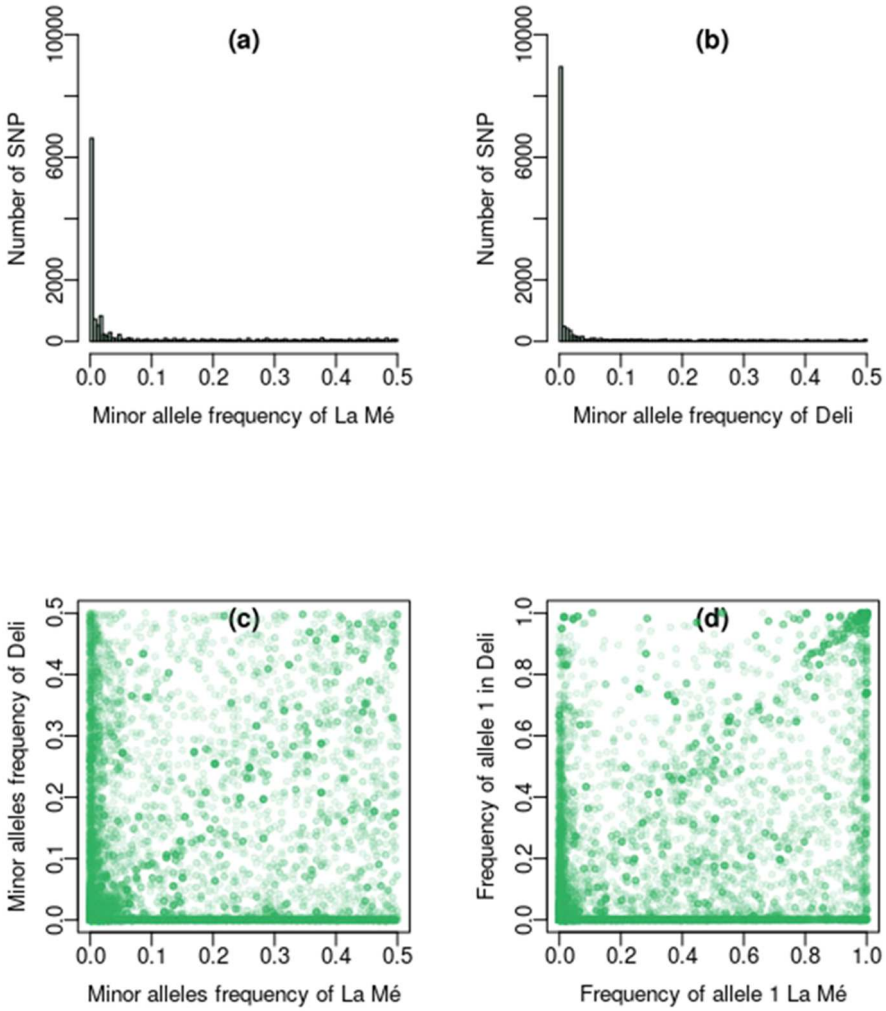


Fig. 1. Imputation and phasing scheme for the production of the SNP datasets used for genomic predictions with the two models PSAM (population-specific effects of SNP alleles model) and ASGM (across-population SNP genotype model). pA, pB, A×B: Deli parents, La Mé parents and Deli×La Mé hybrid ortets, $_{(I)}$ denotes imputed data.



839

840 **Fig. 2**. Distribution of minor allele frequency (MAF) in La Mé (a) and Deli (b) populations, and
841 correlation of MAF (c) and frequency of alternate alleles between La Mé and Deli (d). In (c) and (d)
842 panels, each dot represents an SNP.

843

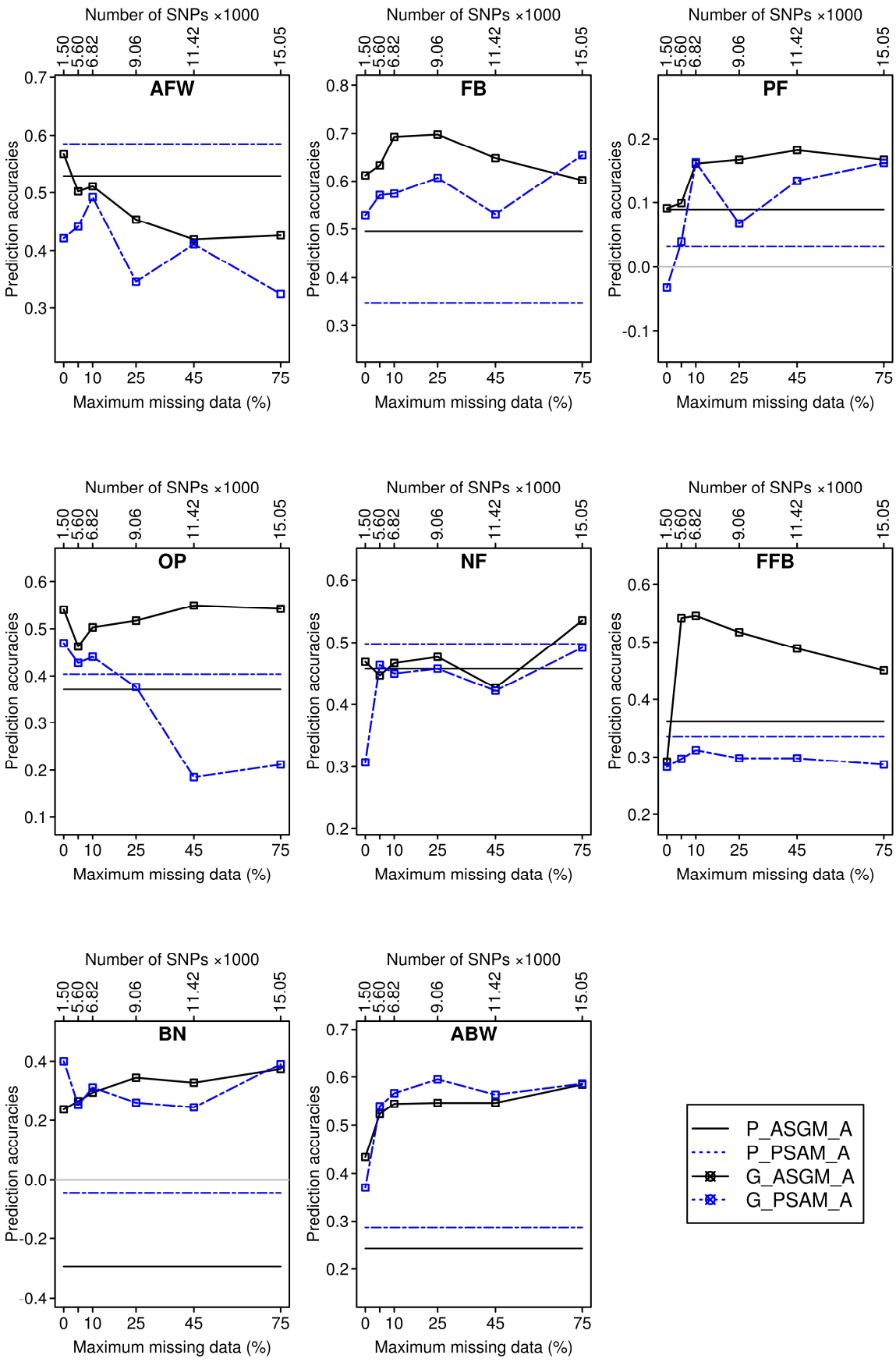

AFW
FB
PF
OP
NF
FFB
BN
ABW
Number of SNPs ×1000
1.50
5.60
6.82
9.06
11.42
15.05
Prediction accuracies
Maximum missing data (%)
0
10
25
45
75
P_ASGM_A
P_PSAM_A
G_ASGM_A
G_PSAM_A

Fig. 3. Prediction accuracies according to traits, SNP datasets and prediction models.

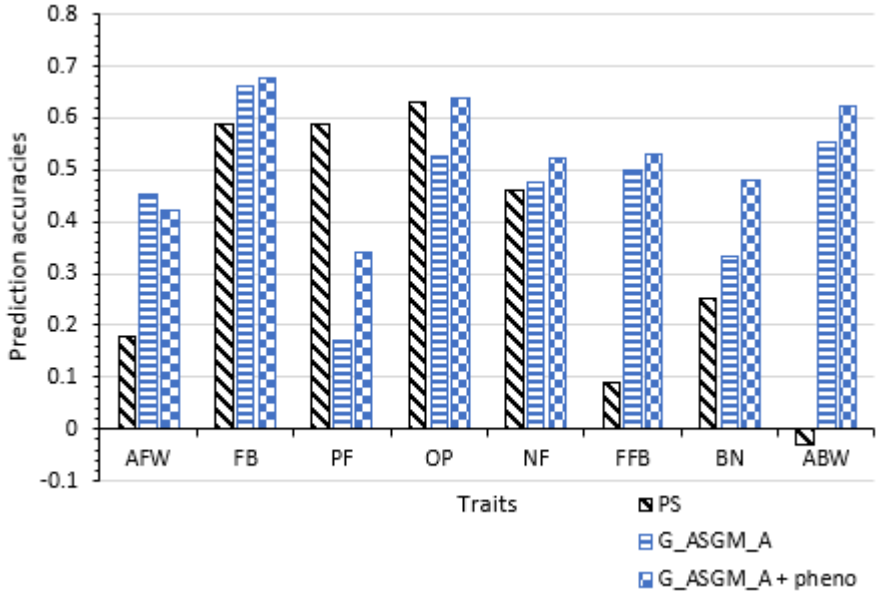
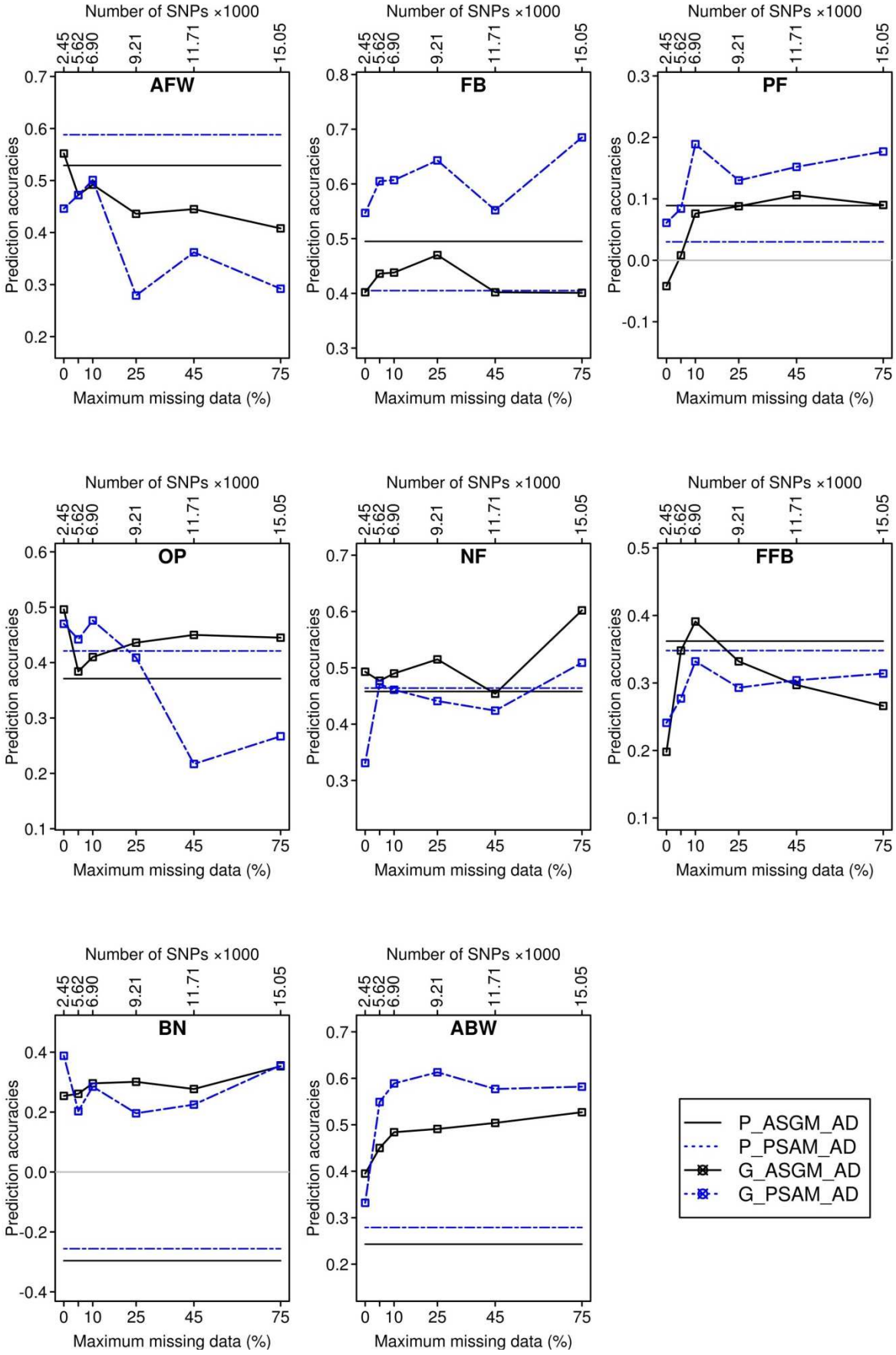


Fig. 4. Prediction accuracies of phenotypic selection (PS) and of the G_ASGM_A model without phenotypic data (G_ASGM_A) and with phenotypic data (G_ASGM_A+pheno) of ortets, on average over the best SNP datasets, and according to trait.

Supplementary data


Number of SNPs ×1000
2.45
5.62
6.90
9.21
11.71
15.05
AFW
FB
PF
OP
NF
FFB
BN
ABW
Prediction accuracies
Maximum missing data (%)
0
10
25
45
75
P_ASGM_AD
P_PSAM_AD
G_ASGM_AD
G_PSAM_AD

855

Supplementary Fig. S. 1 Prediction accuracies according to traits, SNP datasets and prediction models with additive+dominance models.