



**HAL**  
open science

## 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.

### ► To cite this version:

Achille Nyouma, Joseph Martin Bell, Florence Jacob, Virginie Riou, Aurore Manez, et al.. Genomic predictions improve clonal selection in oil palm (*Elaeis guineensis* Jacq.) hybrids. *Plant Science*, 2020, 299, pp.110547. 10.1016/j.plantsci.2020.110547 . hal-02954972

**HAL Id: hal-02954972**

**<https://hal.inrae.fr/hal-02954972v1>**

Submitted on 22 Aug 2022

**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.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 Title

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

3 **Authors and affiliations:**

4 Achille NYOUMA<sup>a,b</sup>, Joseph Martin BELL<sup>a</sup>, Florence JACOB<sup>c</sup>, Virginie RIOU<sup>d,e</sup>, Aurore MANEZ<sup>d,e</sup>,

5 Virginie POMIÈS<sup>d,e</sup>, Leifi NODICHAO<sup>e,f</sup>, Indra SYAHPUTRA<sup>g</sup>, Dadang AFFANDI<sup>g</sup>, Benoit

6 COCHARD<sup>c</sup>, Tristan DURAND-GASSELIN<sup>c</sup>, David CROS<sup>b,d,e\*</sup>

7 <sup>a</sup> Department of Plant Biology, Faculty of Science, University of Yaoundé I, Yaoundé, Cameroon

8 <sup>b</sup> CETIC (African Center of Excellence in Information and Communication Technologies), University

9 of Yaoundé 1, Yaoundé, Cameroon

10 <sup>c</sup> PalmElit SAS, 34980 Montferrier sur Lez, France

11 <sup>d</sup> CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement),

12 UMR AGAP, F-34398 Montpellier, France.

13 <sup>e</sup> University of Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

14 <sup>f</sup> INRAB, CRA-PP, Pobè, Benin

15 <sup>g</sup> P.T. SOCFINDO Medan, Medan, Indonesia

16 \* Corresponding author.

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

18

19

20

21

22

23           **Abstract**

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

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

## 45        1. Introduction

46            The annual yield of palm oil is around four tons per hectare and world production is currently  
47 above 75 million tons of crude palm oil [1]. Most cultivated oil palms (*Elaeis guineensis* Jacq.) are  
48 hybrid cultivars, mainly due to their high yield per hectare. Two parental and heterotic groups are  
49 involved in the production of hybrid cultivars, namely group A, consisting essentially of the Deli  
50 population (Asia) and, to a lesser extent, the Angola population, and group B, involving the other  
51 African breeding populations. Group A produces a small number of large bunches and group B  
52 produces a lot of small bunches. This complementarity and the resulting heterosis expressed on  
53 hybrids through sexual crosses explains why they were widely adopted in the 1960s, leading to a 30%  
54 yield increase [2]. In addition, commercial oil palm material is of *tenera* (*T*) (thin-shelled) fruit type,  
55 resulting from the cross between the thick-shelled *dura* (*D*) of group A and the shell-less and usually  
56 female sterile *pisifera* (*P*) of group B. Selection of hybrids is carried out through progeny tests in a  
57 modified reciprocal recurrent selection (MRRS) breeding scheme [3,4]. The best hybrids are primarily  
58 selected based on the parental general combining abilities (GCA). Although the annual increase of the  
59 oil palm hybrids' yield obtained through genetic improvement reached 1-1.5% over the past decades  
60 [5], this remains insufficient to face the expected increase in the demand.

61            An additional yield increase of 20-30% compared to sexual crosses can be obtained by using  
62 clones (ramets) obtained from the micropropagation of top-ranking commercial hybrid *T* individuals  
63 (ortets) [6]. This allows taking advantage of the within hybrid crosses variability that results from  
64 parental heterozygosity. However, this approach has been hampered for a long time by a floral  
65 epigenetic abnormality producing mantled fruits, which could result in severe production loss. This  
66 abnormality is a somaclonal variation arising during tissue culture due to hypomethylation of the  
67 retrotransposon *Karma* in mantled variants, leading to homeotic transformations and parthenocarpy  
68 [7-9]. The recent understanding of the molecular mechanism involved in the mantled disorder has led  
69 to the possibility of early detection of mantled ramets during the first stages of seedling growth [8],  
70 thus arousing a new impetus for oil palm clonal selection. The evaluation of ortets on their phenotypic  
71 value is possible, but some of the oil palm yield components have a low heritability (e.g. [10] found a

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

83 Genomic selection (GS) [11] is a marker-assisted selection (MAS) method with a high density  
84 of markers on the entire genome, so that at least one marker can be in linkage disequilibrium with each  
85 quantitative trait locus (QTL) [12]. Compared to the previous MAS approach based on QTL detection,  
86 GS takes into account all the markers jointly and without any test of significance. In this way, even  
87 markers capturing small QTL effects are used in the model predicting the genetic values, thus  
88 improving the efficiency of selection. GS is, therefore, the most appropriate MAS method for yield  
89 traits which are usually quantitative, i.e. controlled by many loci with small effect. The GS model is  
90 calibrated (or trained) on individuals genotyped and phenotyped (training set), and predicts the genetic  
91 value of a set of related individuals that are genotyped with the same markers. Before its practical  
92 application, the GS method must be evaluated and the prediction model that gives the highest accuracy  
93 (i.e. the correlation between the predicted and the true genetic values) is retained [13]. The GS  
94 accuracy is estimated in a validation set, made of individuals genotyped and phenotyped and  
95 representative of the population that will be used for application. Oil palm is one of the pioneer  
96 perennial crops on which GS studies have been carried out. The oil palm GS studies provided  
97 prominent results, such as the superiority of GS over both QTL-based MAS and phenotypic selection  
98 [14], and the possibility of increasing the performance of sexual hybrid crosses by genomic  
99 preselection before progeny-tests [15]. The main advantages of GS for the oil palm are its ability to

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

107         Given that ortets come from a cross between two oil palm origins, the genomic prediction of  
108 their genetic values can be done by two modeling approaches [20], which are the genomic extensions  
109 of the modeling approach developed by Stuber and Cockerham [21] for interpopulation hybrids. The  
110 first one, the population-specific effects of single nucleotide polymorphism (SNP) alleles model  
111 (PSAM, or BSAM in the animal breeding literature, for breed instead of population), considers that  
112 alleles of the same marker have different effects in the hybrids depending on their population of origin,  
113 whereas the second approach, the across-population SNP genotype model (ASGM), considers that  
114 alleles of a marker have the same effect regardless of their population of origin. Studies in livestock  
115 showed that BSAM can outperform ASGM in terms of accuracy with a low number of SNPs, a large  
116 training set and slightly related or unrelated individuals [20]. However, to our knowledge, in the  
117 context of plant hybrids, these types of models were only compared in simulated maize populations  
118 [22].

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

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

131

## 132 **2. Materials and methods**

### 133 *2.1. Plant materials and experimental designs*

134 The plant material used to train the GS model comes from controlled crosses between Deli and  
135 La Mé (LM) individuals. Deli material comes from four ancestors of an unknown area of Africa  
136 planted in Indonesia in 1848. The La Mé material used here comes from three founders collected in  
137 Ivory Coast between 1924 and 1930 [15,23]. For bunch production predictions, the training set was  
138 composed of 295 progeny-test crosses planted from 1995 to 2000 at Aek Loba Timur (ALT) and  
139 involving 108 Deli and 102 La Mé. For bunch quality predictions, a sample of 279 crosses involving  
140 103 Deli and 100 La Mé parents were used (Table 1). The pedigrees of these populations are known  
141 over several generations (see Cros et al. [12]). ALT is located at 2° 39' N – 99° 42' E in North  
142 Sumatra, on the SOCFINDO estate (Indonesia) and is constituted of 28 trials planted on deep loamy  
143 sand soils, with low water deficit and high insolation, and benefiting from standard cultural practices  
144 [24]. The experimental design used in these trials was either a balanced lattice of four to five ranks or  
145 randomized complete block designs (RCBD), described in detail by Cros et al. [15].

146 The validation set was composed of 42 Deli × La Mé *tenera* ortets, evaluated in clonal trials  
147 involving on average 69 ramets per clone for production traits and a subset of 34 ramets per clone for  
148 quality traits. The ramets were established in three out of the 28 trials of ALT and were planted in  
149 1995 and 1998 (Table 1). The 42 ortets were chosen among individuals from various hybrid crosses  
150 planted on seven trials of an earlier set of progeny tests, located at Aek Kwasan 1 (AK1), which was  
151 also located on the SOCFINDO estate and benefited from the same agricultural practices. The  
152 plantation of the seven trials of AK1 took place between 1975 and 1979. The 42 ortets come from 17  
153 families of full sibs with 16 La Mé parents and 12 Deli parents. These families were composed of one  
154 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  $\sigma$  the standard deviation and  
167  $\mu$  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



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

184

#### 185 *2.4. Imputation of missing SNP data and phasing*

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

199

#### 200 *2.5. Definition of SNP datasets*

201 To quantify how the characteristics of the SNP dataset (i.e. maximum percentage of missing  
202 data allowed per SNP,  $p_{max}$ , and resulting number of SNPs,  $n_{snp}$ ) affected the GS accuracy, we made  
203 genomic predictions using different SNP datasets with varying maximum percentage of missing data  
204 per SNP, as shown in table 2. Thereby, for the rest of the study, the SNP dataset will refer to an SNP  
205 matrix with a given number of SNPs resulting from the filtering made on the maximum percentage of  
206 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.

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

236

### 237 2.6.1. Across-population SNP genotype models (ASGM)

238 The model used for the G\_ASGM\_AD approach was as follows:

$$239 \quad y = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{g}_i + \mathbf{Z}_2\mathbf{g}_{Deli \times LM} + \mathbf{Z}_3\mathbf{b} + \mathbf{Z}_4\mathbf{p} + \varepsilon$$

240 with:  $y$  the observed phenotypes of the training hybrid individuals,  $\beta$  the vector of fixed effects

241 (phenotypic mean, trial effects, block effects and, for bunch production traits, age),

242  $\mathbf{g}_i \sim \text{N}(0, \mathbf{H}_i\sigma_{a_i}^2)$  the individual additive genetic effects,  $\mathbf{g}_{Deli \times LM} \sim \text{N}(0, \mathbf{H}_{Deli \times LM}\sigma_{d_{Deli \times LM}}^2)$  the

243 genetic dominance effects,  $\mathbf{b} \sim \text{N}(0, \mathbf{I}\sigma_b^2)$  the incomplete block effect, and  $\mathbf{p} \sim \text{N}(0, \mathbf{I}\sigma_p^2)$  the

244 elementary plot effects.  $\mathbf{X}$ ,  $\mathbf{Z}_1$ ,  $\mathbf{Z}_2$ ,  $\mathbf{Z}_3$  and  $\mathbf{Z}_4$  are the incidence matrices associated to  $\beta$ ,  $\mathbf{g}_i$ ,

245  $\mathbf{g}_{Deli \times LM}$ ,  $\mathbf{b}$  and  $\mathbf{p}$  respectively.  $\mathbf{H}_i\sigma_{a_i}^2$  and  $\mathbf{H}_{Deli \times LM}\sigma_{d_{Deli \times LM}}^2$  are the variance-covariance matrices

246 associated with  $\mathbf{g}_i$  and  $\mathbf{g}_{Deli \times LM}$ , respectively.  $\sigma_{a_i}^2$  and  $\sigma_{d_{Deli \times LM}}^2$  are the additive and dominance

247 variances, respectively.  $\varepsilon \sim \text{N}(0, \mathbf{I}\sigma_\varepsilon^2)$  is the vector of residual effects and  $\mathbf{I}$  the identity matrix. To

248 implement this model in practice, two specificities of our dataset had to be taken into account. First, a

249 few parents of the training crosses were not genotyped (Table 1), and the  $\mathbf{H}_i$  matrices had therefore to

250 be made with the genealogical data of hybrid crosses with ungenotyped parents and with the SNP data

251 of hybrid crosses with genotyped parents (computed with the SNP data of their parents, see below)

252 and of the ortets. All  $\mathbf{H}_i$  matrices subsequently in this paper will refer to matrices combining

253 genealogical and genomic information.  $\mathbf{H}_i^{-1}$  is the inverse of  $\mathbf{H}_i$ , computed according to Miszta et al.

254 [33] as:  $\mathbf{H}_i^{-1} = \mathbf{A}_i^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}_i^{-1} - \mathbf{A}_{i22}^{-1} \end{bmatrix}$ , where  $\mathbf{G}_i^{-1}$  and  $\mathbf{A}_{i22}^{-1}$  are the inverse of the realized and the

255 genealogical additive relationship matrices, respectively, of the 42 ortets and the hybrid crosses with

256 genotyped parents, and  $\mathbf{A}_i^{-1}$  is the inverse of the genealogical relationship matrix of all hybrid crosses

257 (i.e. the few with ungenotyped parents and the ones with genotyped parents) and the 42 ortets. Second,

258 the phenotyped individuals constituting the hybrid crosses were not genotyped while they had to be

259 connected to the validation ortets through their genomic relationships (only the parents of the hybrids  
260 were genotyped, except a few parents that were not genotyped and for which the genealogical  
261 relationships were used, as explained above). To get genotypes for the hybrid crosses with genotyped  
262 parents, we computed for each cross the mean genotypes expected from the parental genotypes (i.e. for  
263 SNP  $j$  in cross  $i$ , the mean number of copies of the minor allele of SNP  $j$  expected to be found in the  
264 hybrid individuals of  $i$ ), assuming this was relevant considering the relatively large number of  
265 individuals per cross (Table 1). The genomic additive relationship matrix  $\mathbf{G}$  was obtained as:  $\mathbf{G} =$   
266  $\frac{\mathbf{X}'\mathbf{X}}{2\sum_{l=1}^{n_{\text{SNP}}} p_l(1-p_l)}$ , with  $\mathbf{X} = \mathbf{Z} - \mathbf{P}$ ,  $\mathbf{X}'$  the transpose of matrix  $\mathbf{X}$ ,  $\mathbf{Z}$  the SNP matrix containing the  
267 number of copies of the minor allele at an SNP (ranging from 0 to 2),  $\mathbf{P}$  a matrix given by  $\mathbf{P} = 2(p_l -$   
268  $0.5)$ , and  $p_l$  the frequency of the minor allele at SNP  $l$  [34].  $\mathbf{H}_{\text{Deli} \times \text{LM}}$  is the dominance relationship  
269 matrix combining genomic dominance relationships between crosses with parents and clones, and  
270 genealogical dominance relationships between the few crosses with ungenotyped parents.  
271  $\mathbf{H}_{\text{Deli} \times \text{LM}}^{-1}$  was computed following the same method as  $\mathbf{H}_i^{-1}$  except that the additive relationship  
272 matrices were replaced by the dominance relationship matrices. The realized dominance relationship  
273 matrix  $\mathbf{G}_D$  was computed according to Su et al. [35] as:  $\mathbf{G}_D = \frac{\mathbf{\Pi}\mathbf{\Pi}'}{2\sum p_l q_l (1-2p_l q_l)}$ , with  $\mathbf{\Pi}$  the  $n \times m$   
274 matrix ( $n$ : number of hybrid crosses and clones and  $m$ : number of SNPs) of heterozygosity coefficients  
275 with element  $\mathbf{\Pi}_{kl} = 0 - p_l q_l$  if clone or ortet  $k$  is homozygous and  $\mathbf{\Pi}_{kl} = 1 - p_l q_l$  if it is  
276 heterozygous at locus  $l$ , and  $p_l$  and  $q_l$  the frequencies of the first and the second allele at locus  $l$ . The  
277 purely additive approach ASGM\_A used the same model without the dominance effect.

278 For the P\_ASGM\_A and P\_ASGM\_AD,  $\mathbf{H}_i$  was replaced by the additive genealogical  
279 relationship matrix  $\mathbf{A}_i$  and, for P\_ASGM\_AD,  $\mathbf{H}_{\text{Deli} \times \text{LM}}$  was replaced by the genealogical dominance  
280 relationship matrix.

281 The estimated genetic value for the validation clones was  $\hat{g}_i$  and, for G\_ASGM\_AD and  
282 P\_ASGM\_AD,  $\hat{g}_i + \hat{g}_{\text{Deli} \times \text{LM}}$ .

283

284 2.6.2. Population-specific effects of SNP alleles models (PSAM)

285 The model used for G\_PSAM\_AD was as follows:

286 
$$y = X\beta + Z_1g_{Deli} + Z_2g_{LM} + Z_3g_{Deli \times LM} + Z_4b + Z_5p + \varepsilon$$

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

306 For P\_PSAM\_A and P\_PSAM\_AD,  $H_{Deli}$  and  $H_{LM}$  were replaced by the additive  
 307 genealogical relationship matrices  $A_{Deli}$  and  $A_{LM}$  and, for P\_PSAM\_AD,  $H_{Deli \times LM}$  was replaced by  
 308 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

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

340

### 341 **3. Results**

#### 342 *3.1. Distribution of frequencies of minor and alternate alleles across population*

343 The distribution of MAF in both Deli and La Mé populations showed a reduction in the  
344 number of SNPs with the increase of MAF (Fig. 2). The MAF ranged from 0 to 0.5 for both La Mé  
345 and Deli populations and the average was 0.1 for La Mé (Fig. 2a) and 0.07 for Deli (Fig. 2b). Most  
346 SNPs had low MAF values ( $<0.05$ ) in both populations. La Mé populations had 65.6% SNPs with  
347  $MAF < 0.05$ , against 73.3% SNPs in Deli (i.e. 11.7% more SNPs with low MAF in Deli). In contrast,  
348 fewer SNPs had high MAF ( $>0.40$ ) in both populations, and they were higher in proportion in La Mé  
349 (8.2% SNPs) than in Deli (4.8%). This showed the lower genetic diversity of Deli parents compared to  
350 La Mé, which resulted from their contrasted history with more generations of selection, drift and  
351 inbreeding in Deli than in La Mé.

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

361 Deli population. Also, the number of SNPs segregating with  $MAF > 0.05$  in both populations was low  
362 (14.8% of all SNPs). Despite these differences, a large number of SNPs (53.7% of all SNPs) had  
363  $MAF < 0.05$  in both populations, showing segregation with rare alleles in both Deli and La Mé.  
364 However, correlation of the frequency of the alternate allele between La Mé and Deli (Fig. 2d) over all  
365 SNPs showed that 62.8% of SNPs have a frequency of alternate allele smaller than 0.05 in one  
366 population and greater than 0.95 in the other population, i.e. fixed or almost fixed in the two  
367 populations but for different alleles. Hence, given that most of the SNPs (85.2%) have either  
368  $MAF < 0.05$  in one population and  $MAF \geq 0.05$  in the other population (31.5%), or  $MAF < 0.05$  in both  
369 populations but for different alleles (53.7%), the use of PSAM is justified.

370

### 371 *3.2. Effect of GS prediction model and SNP dataset on prediction accuracy*

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

377 On average over traits and SNP datasets, G\_ASGM\_A was more accurate (0.45) than  
378 G\_PSAM\_A (0.37), with the mean prediction accuracy per trait over SNP datasets ranging from 0.14  
379 (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  
380 obtained a mean prediction accuracy greater than G\_PSAM\_A for six traits out of eight, with  
381 G\_PSAM\_A being on average slightly more accurate than G\_ASGM\_A for ABW and equal for BN  
382 (Table 3). Considering the maximum accuracy over all SNP datasets, the prediction accuracy ranged  
383 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  
384 3), and, here, G\_PSAM\_A was more accurate for both BN and ABW, although slightly. Considering  
385 the different SNP datasets and traits, large differences in prediction accuracy between G\_ASGM\_A  
386 and G\_PSAM\_A were observed, up to +0.36 in favour of G\_ASGM\_A with OP at  $p_{max} = 45\% - n_{SNP} =$   
387 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

416 plateauing (except for AFW) and slightly decreasing in some cases. This suggested that more useful  
417 information was captured for prediction purposes when using more SNPs (to a certain limit) and,  
418 again, that the percentage of missing data was of lesser importance. On the other hand, a reduction of  
419 accuracies was observed with SNP density for AFW. For G\_PSAM\_A, prediction accuracies  
420 increased, and usually plateaued, for only four traits (FB, PF, NF and ABW). For the other traits,  
421 prediction accuracies remained stable or tended to decrease with increasing marker density and  
422 maximum percentage of missing SNP data. Thus, the accuracy of OP, for instance, decreased around  
423 59.6% from  $p_{max}=0\%-n_{snp}=1,497$  to  $p_{max}=45\%-n_{snp}=11,425$  (Fig. 3).

### 424 3.3. Comparison of prediction accuracies of PS and GS

425 Figure 4 presents the prediction accuracies of PS and the mean prediction accuracy of  
426 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),  
427 with (G\_ASGM\_A+pheno) and without phenotypic data of the ortets. Variation of PS accuracy was  
428 large between traits, going from -0.03 for ABW to 0.63 for OP. Very low PS accuracies (<0.1) were  
429 obtained for ABW and FFB, meaning that PS would have been inefficient for these two traits. The  
430 highest PS accuracies were achieved in OP (0.63) and PF (0.59) (Table 6 and Fig. 4). These two traits  
431 are known to have moderate to high heritability in the oil palm [2] and are consequently routinely used  
432 for preselection before clonal trials. This was the case here, as indicated by the intensity of PS for  
433 these two traits, which was the highest among the eight traits studied (Table 6).

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

442 against -0.03. PS only surpassed G\_ASGM\_A for two traits (PF and OP) and G\_ASM\_A+pheno for  
443 one trait (PF).

444

#### 445 **4. Discussion**

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

452

##### 453 *4.1. Improving the genetic progress of clonal breeding with GS*

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

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

468 clonal trials is usually applied to PF and OP, as they have the highest  $H^2$  [39]. In our dataset, these  
469 traits indeed had high  $H^2$ , with PS prediction accuracy  $>0.5$  (Fig. 4) (although it was not clear why FB  
470 had a similar  $H^2$ , while it is usually among the traits with low  $H^2$ ). Therefore, considering that breeders  
471 use 0.5 as the minimum prediction accuracy for applying PS before clonal trials, they would now  
472 apply GS to four traits (FB, OP, FFB and ABW) (Fig. 4), with a similar mean prediction accuracy over  
473 these traits with GS (0.56) compared to PS (0.60 over FB, PF and OP). Interestingly, the two traits that  
474 had a prediction accuracy lower with G\_ASGM\_A than with PS, i.e. PF and OP, were the ones for  
475 which the 42 ortets were submitted to the strongest phenotypic selection before clonal trials. In  
476 particular, PF had the highest intensity of phenotypic selection (0.68) and also had much lower  
477 prediction accuracy with G\_ASGM\_A than with PS. We hypothesized this occurred as the phenotypic  
478 preselection led to the fixation of many genes controlling these traits, and in particular PF, in the 42  
479 ortets, thus making that the relationships computed over the genome-wide SNPs no longer matched  
480 with the relationships at the genes. This hypothesis could be investigated using a validation set that  
481 was not submitted to phenotypic preselection. Such a study would be of great interest as, in case our  
482 hypothesis could be confirmed, the breeders would likely get in practice a higher GS accuracy for PF  
483 and OP, as the seedlings comprising the population of application would not be preselected. In this  
484 case, GS before the clonal trials would be even more useful. Second, a GS-based approach would also  
485 increase the genetic progress by higher selection intensity compared to PS: GS would be applied to  
486 nursery individuals, i.e. possibly in the thousands, while PS is currently applied to the small number of  
487 individuals planted in the progeny tests trials (i.e. normally 10 to 50 per cross) [9]. Third, making the  
488 selection in the best possible crosses instead of the best crosses evaluated would be an improvement in  
489 terms of genetic progress, as the best possible crosses were likely not present in the progeny tests, due  
490 to the high degree of incompleteness of the mating designs. It is also possible to make these crosses in  
491 the context of phenotypic clonal selection, but in this case, the selection process would require around  
492 10 more years of phenotypic evaluations in these elite crosses to identify the candidate ortets for the  
493 clonal trials [16].

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

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

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

509

#### 510 4.2. Effects of prediction model and SNP dataset on prediction accuracies

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

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

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

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

552 megabase pair, Mbp) than with high marker density (10 SNP per Mbp). Here, considering the length  
553 of the oil palm genome is 1.8 Gb [44], the investigated range of SNP density was similar, going from  
554 0.8 to 8.4 SNP per Mbp. Moreover, Lopes et al. [45] obtained similar prediction accuracies between  
555 G\_ASGM\_A and G\_PSAM\_A with high SNP density (31,930 SNPs). We did not find SNP density to  
556 have such an effect on the prediction accuracy of G\_PSAM\_A or on the relative performance of  
557 G\_PSAM\_A and G\_ASGM\_A. This likely results from the fact that, in our study, SNP density varied  
558 with SNP quality, with higher SNP numbers meaning a higher percentage of missing data. These two  
559 parameters, therefore, seem to interact on the prediction accuracy of the two models investigated.  
560 However, the fact that the mean GS accuracy over the traits increased with the number of SNPs and  
561 plateaued from 6,898 SNPs indicated that SNP density was of greater importance for the prediction  
562 accuracy than the percentage of missing data per SNP.

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

571 GS prediction models (G\_ASGM\_A and G\_PSAM\_A) were usually more accurate than their  
572 respective control pedigree-based models (P\_ASGM\_A and P\_PSAM\_A). The superiority of GS  
573 models shows that, even for unobserved individuals, GS models can account for both Mendelian  
574 sampling terms of siblings in a family and for family effects, while pedigree-based models can only  
575 account, at best, for family effects, as already found in previous oil palm GS studies [16].

576 However, G\_ASGM\_A outperformed its control pedigree-based model more often than  
577 G\_PSAM\_A. Thus, G\_PSAM\_A remained less accurate than P\_PSAM\_A for all the SNP datasets in  
578 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

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

639

#### 640 **Data availability**

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

643

#### 644 **Conflict of interests**

645 The authors declare no conflict of interest.

646

#### 647 **Author contributions**

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

652

- 654 [1] USDA, <http://www.fas.usda.gov/data/oilseeds-world-markets-and-trade>. Accessed January 13,  
655 2020., 2020.
- 656 [3] J.P. Gascon, C. Berchoux, Caractéristique de la production d'*Elaeis guineensis* (Jacq.) de  
657 diverses origines et de leurs croisements - Application à la sélection du palmier à huile,  
658 Oléagineux. 19 (1964) 75–84.
- 659 [4] J. Meunier, J. Gascon, Le schéma général d'amélioration du palmier à huile à l'IRHO,  
660 Oléagineux. 27 (1972) 1–12.
- 661 [5] A. Rival, P. Levang, Palms of controversies: oil palm and development challenges, CIFOR,  
662 Jakarta, Indonésie, 2014. [http://www.cifor.org/publications/pdf\\_files/Books/BLevang1401.pdf](http://www.cifor.org/publications/pdf_files/Books/BLevang1401.pdf)  
663 (accessed October 23, 2014).
- 664 [6] R. Corley, I. Law, The future for oil palm clones, in: 1997: pp. 279–289.
- 665 [7] E. Jaligot, A. Rival, T. Beulé, S. Dussert, J.-L. Verdeil, Somaclonal variation in oil palm (*Elaeis*  
666 *guineensis* Jacq.): the DNA methylation hypothesis, *Plant Cell Reports*. 19 (2000) 684–690.
- 667 [8] M. Ong-Abdullah, J.M. Ordway, N. Jiang, S. Ooi, S.-Y. Kok, N. Sarpan, N. Azimi, A.T.  
668 Hashim, Z. Ishak, S.K. Rosli, Loss of Karma transposon methylation underlies the mantled  
669 somaclonal variant of oil palm, *Nature*. 525 (2015) 533.
- 670 [9] A.C. Soh, S. Mayes, J.A. Roberts, *Oil Palm Breeding: Genetics and Genomics*, CRC Press,  
671 2017.
- 672 [10] B. Nouy, J.-C. Jacquemard, E. Suryana, F. Potier, K. Konan, T. Durand-Gasselín, The expected  
673 and observed characteristics of several oil palm (# *Elaeis guineensis*# Jacq.) clones, (2006).
- 674 [11] T.H. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-  
675 wide dense marker maps, *Genetics*. 157 (2001) 1819–1829.
- 676 [12] M.E. Goddard, B.J. Hayes, Genomic selection, *J. Anim. Breed. Genet.* 124 (2007).  
677 <https://doi.org/10.1111/j.1439-0388.2007.00702.x>.
- 678 [13] D. Grattapaglia, O.B. Silva-Junior, R.T. Resende, E.P. Cappa, B.S. Müller, B. Tan, F. Isik, B.  
679 Ratcliffe, Y.A. El-Kassaby, Quantitative genetics and genomics converge to accelerate forest  
680 tree breeding, *Frontiers in Plant Science*. 9 (2018) 1693.
- 681 [14] C. Wong, R. Bernardo, Genomewide selection in oil palm: increasing selection gain per unit  
682 time and cost with small populations, *Theoretical and Applied Genetics*. 116 (2008) 815–824.
- 683 [15] D. Cros, S. Bocs, V. Riou, E. Ortega-Abboud, S. Tisné, X. Argout, V. Pomiès, L. Nodichao, Z.  
684 Lubis, B. Cochard, Genomic preselection with genotyping-by-sequencing increases performance  
685 of commercial oil palm hybrid crosses, *BMC Genomics*. 18 (2017) 839.  
686 <https://doi.org/10.1186/s12864-017-4179-3>.
- 687 [16] A. Nyouma, J.M. Bell, F. Jacob, D. Cros, From mass selection to genomic selection: one century  
688 of breeding for quantitative yield components of oil palm (*Elaeis guineensis* Jacq.), *Tree*  
689 *Genetics & Genomes*. 15 (2019) 69. <https://doi.org/10.1007/s11295-019-1373-2>.
- 690 [17] Q.B. Kwong, A.L. Ong, C.K. Teh, F.T. Chew, M. Tammi, S. Mayes, H. Kulaveerasingam, S.H.  
691 Yeoh, J.A. Harikrishna, D.R. Appleton, Genomic selection in commercial perennial crops:  
692 applicability and improvement in oil palm (*Elaeis Guineensis* Jacq.), *Scientific Reports*. 7 (2017)  
693 2872.
- 694 [18] R. Durán, F. Isik, J. Zapata-Valenzuela, C. Balocchi, S. Valenzuela, Genomic predictions of  
695 breeding values in a cloned *Eucalyptus globulus* population in Chile, *Tree Genetics & Genomes*.  
696 13 (2017) 74.
- 697 [19] D. Cros, L. Mbo-Nkoulou, J.M. Bell, J. Oum, A. Masson, M. Soumahoro, D.M. Tran, Z.  
698 Achour, V. Le Guen, A. Clement-Demange, Within-family genomic selection in rubber tree  
699 (*Hevea brasiliensis*) increases genetic gain for rubber production, *Industrial Crops and Products*.  
700 138 (2019) 111464. <https://doi.org/10.1016/j.indcrop.2019.111464>.
- 701 [20] N. Ibáñez-Escriche, R. Fernando, A. Toosi, J. Dekkers, Genomic selection of purebreds for  
702 crossbred performance, *Genetics Selection Evolution*. 41 (2009) 12.
- 703 [21] C. Stuber, C.C. Cockerham, Gene effects and variances in hybrid populations, *Genetics*. 54  
704 (1966) 1279.

- 705 [22] F. Technow, C. Riedelsheimer, Tobias A. Schrag, Albrecht E. Melchinger, Genomic prediction of  
706 hybrid performance in maize with models incorporating dominance and population specific  
707 marker effects, *Theor Appl Genet.* 125 (2012) 1181–1194. <https://doi.org/10.1007/s00122-012-1905-8>.  
708
- 709 Corley, R.H.V., Tinker, P.B., 2016. *The oil palm*, 5th ed. Wiley-Blackwell, Chichester, UK.  
710 <https://doi.org/10.1002/9781118953297>
- 711 [24] F. Potier, B. Nouy, A. Flori, J. Jacquarmard, H. Edyana Suryana, T. Durand-Gasselin, Yield  
712 potential of oil palm (*Elaeis guineensis* Jacq) clones: preliminary results observed in the Aek  
713 Loba genetic block in Indonesia, in: 2006.
- 714 [25] J. He, X. Zhao, A. Laroche, Z.-X. Lu, H. Liu, Z. Li, Genotyping-by-sequencing (GBS), an  
715 ultimate marker-assisted selection (MAS) tool to accelerate plant breeding, *Frontiers in Plant*  
716 *Science.* 5 (2014) 484. <https://doi.org/10.3389/fpls.2014.00484>.
- 717 [26] R.J. Elshire, J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, S.E. Mitchell, A  
718 robust, simple genotyping-by-sequencing (GBS) approach for high diversity species, *PLoS One.*  
719 6 (2011) e19379.
- 720 [27] J.C. Glaubitz, T.M. Casstevens, F. Lu, J. Harriman, R.J. Elshire, Q. Sun, E.S. Buckler, TASSEL-  
721 GBS: a high capacity genotyping by sequencing analysis pipeline, *PloS One.* 9 (2014) e90346.
- 722 [28] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, *Nature Methods.* 9  
723 (2012) 357.
- 724 [29] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G.  
725 Lunter, G.T. Marth, S.T. Sherry, The variant call format and VCFtools, *Bioinformatics.* 27  
726 (2011) 2156–2158.
- 727 [30] S.R. Browning, B.L. Browning, Rapid and accurate haplotype phasing and missing-data  
728 inference for whole-genome association studies by use of localized haplotype clustering, *The*  
729 *American Journal of Human Genetics.* 81 (2007) 1084–1097.
- 730 [31] S.A. Clark, J. van der Werf, Genomic best linear unbiased prediction (gBLUP) for the estimation  
731 of genomic breeding values, in: *Genome-Wide Association Studies and Genomic Prediction*,  
732 Springer, 2013: pp. 321–330.
- 733 [32] D. Habier, R. Fernando, J. Dekkers, The impact of genetic relationship information on genome-  
734 assisted breeding values, *Genetics.* 177 (2007) 2389–2397.
- 735 [33] I. Misztal, I. Aguilar, D. Johnson, A. Legarra, S. Tsuruta, T. Lawlor, A unified approach to  
736 utilize phenotypic, full pedigree and genomic information for a genetic evaluation of Holstein  
737 final score, *Interbull Bulletin.* (2009) 240.
- 738 [34] P.M. VanRaden, Efficient methods to compute genomic predictions, *Journal of Dairy Science.*  
739 91 (2008) 4414–4423.
- 740 [35] G. Su, O.F. Christensen, T. Ostersen, M. Henryon, M.S. Lund, Estimating additive and non-  
741 additive genetic variances and predicting genetic merits using genome-wide dense single  
742 nucleotide polymorphism markers, *PloS One.* 7 (2012) e45293.
- 743 [36] J.H. Steiger, Tests for comparing elements of a correlation matrix., *Psychological Bulletin.* 87  
744 (1980) 245.
- 745 [37] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*,  
746 Northwestern University, Evanston, Illinois, 2018. <https://CRAN.R-project.org/package=psych>.
- 747 [38] D. Cros, B. Tchounke, L. Nkague-Nkamba, Training genomic selection models across several  
748 breeding cycles increases genetic gain in oil palm in silico study, *Molecular Breeding.* 38 (2018)  
749 89. <https://doi.org/10.1007/s11032-018-0850-x>.
- 750 [39] R.H.V. Corley, P.B. Tinker, *Vegetative Propagation and Biotechnology*, in: *The Oil Palm*, John  
751 Wiley & Sons, Ltd, 2016: pp. 208–224. <https://doi.org/10.1002/9781118953297.ch7>.
- 752 [40] S. Tisné, M. Denis, D. Cros, V. Pomiès, V. Riou, I. Syahputra, A. Omoré, T. Durand-Gasselin,  
753 J.-M. Bouvet, B. Cochard, Mixed model approach for IBD-based QTL mapping in a complex oil  
754 palm pedigree, *BMC Genomics.* 16 (2015) 798.
- 755 [41] F. Technow, T.A. Schrag, W. Schipprack, E. Bauer, H. Simianer, A.E. Melchinger, Genome  
756 Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program  
757 of Maize, *Genetics.* 197 (2014) 1343. <https://doi.org/10.1534/genetics.114.165860>.

- 758 [42] J. Zeng, A. Toosi, R.L. Fernando, J.C. Dekkers, D.J. Garrick, Genomic selection of purebred  
759 animals for crossbred performance in the presence of dominant gene action, *Genetics Selection*  
760 *Evolution*. 45 (2013) 11.
- 761 [43] H. Esfandyari, A.C. Sørensen, P. Bijma, A crossbred reference population can improve the  
762 response to genomic selection for crossbred performance, *Genetics Selection Evolution*. 47  
763 (2015) 76.
- 764 [44] R. Singh, M. Ong-Abdullah, E.-T.L. Low, M.A.A. Manaf, R. Rosli, R. Nookiah, L.C.-L. Ooi, S.  
765 Ooi, K.-L. Chan, M.A. Halim, Oil palm genome sequence reveals divergence of interfertile  
766 species in Old and New worlds, *Nature*. 500 (2013) 335.
- 767 [45] M.S. Lopes, H. Bovenhuis, A.M. Hidalgo, J.A. Van Arendonk, E.F. Knol, J.W. Bastiaansen,  
768 Genomic selection for crossbred performance accounting for breed-specific effects, *Genetics*  
769 *Selection Evolution*. 49 (2017) 51.
- 770 [46] J.-F. Zuo, Y. Niu, P. Cheng, J.-Y. Feng, S.-F. Han, Y.-H. Zhang, G. Shu, Y. Wang, Y.-M.  
771 Zhang, Effect of marker segregation distortion on high density linkage map construction and  
772 QTL mapping in Soybean (*Glycine max* L.), *Heredity*. (2019). [https://doi.org/10.1038/s41437-](https://doi.org/10.1038/s41437-019-0238-7)  
773 [019-0238-7](https://doi.org/10.1038/s41437-019-0238-7).
- 774 [47] C. Li, G. Bai, S. Chao, Z. Wang, A high-density SNP and SSR consensus map reveals  
775 segregation distortion regions in wheat, *BioMed Research International*. 2015 (2015).
- 776 [48] J.C. Reif, F.-M. Gumpert, S. Fischer, A.E. Melchinger, Impact of interpopulation divergence on  
777 additive and dominance variance in hybrid populations, *Genetics*. 176 (2007) 1931–1934.
- 778

779 **Tables**780 **Table 1**

781 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
<b>Number of crosses or ortets</b>	295	279	42	42
<b>Number of individuals or ramets</b>	19,668	12,341	2,908	1,439
<b>Average number of individuals per cross or ramets per clone (min–max)</b>	67 (17-503)	44 (21-274)	69 (5-138)	34 (4-74)
<b>Number of Deli parents (genotyped)</b>	108 (93)	103 (90)	16	16
<b>Number of La Mé parents (genotyped)</b>	102 (91)	100 (89)	12	12
<b>Age at time of data collection (years)</b>	3-7	5-9	3-7	5-9

782

783 **Table 2**

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

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

786

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



797 **Table 4**

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

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

804

805 **Table 5**

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

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

---

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

---

813 **Table 6**

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

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

815

816 **Figures**

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

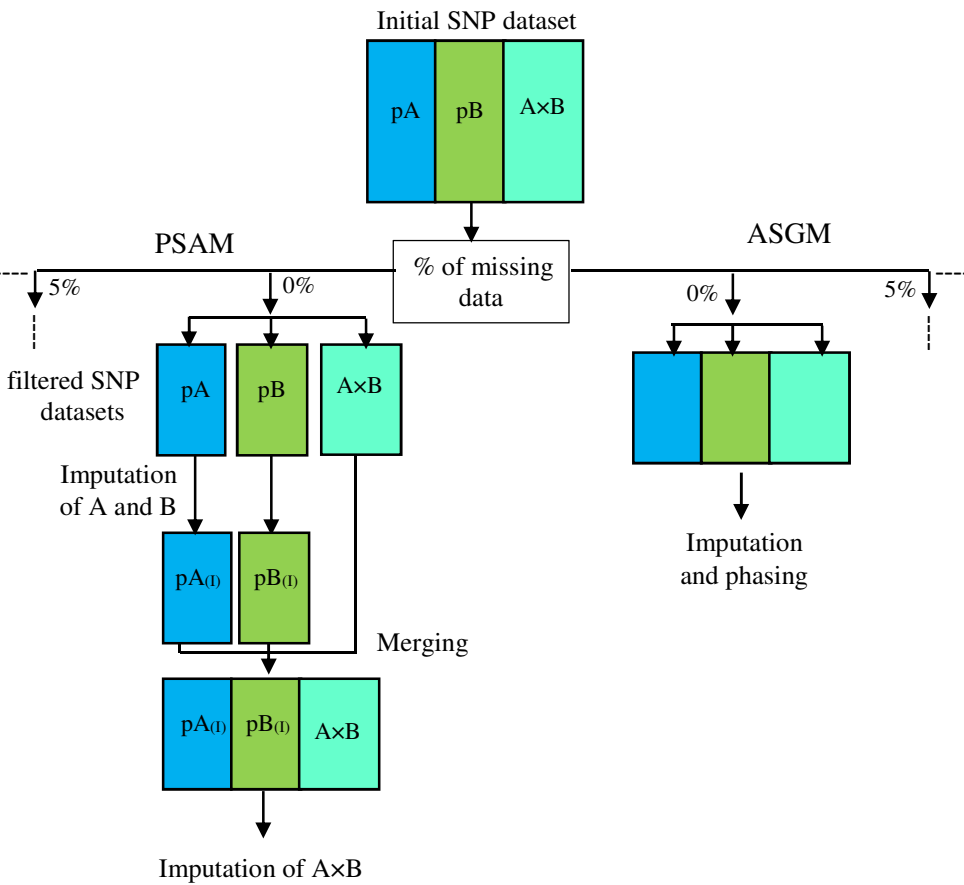
834

835

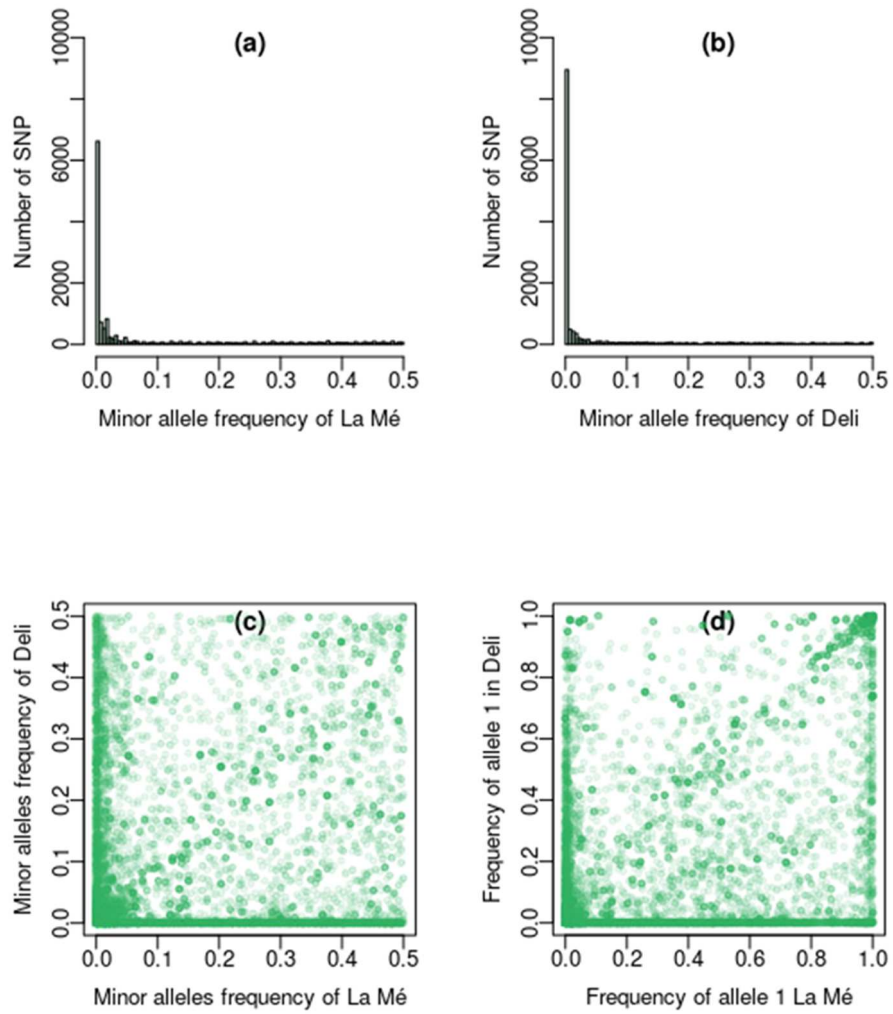
836

837

838



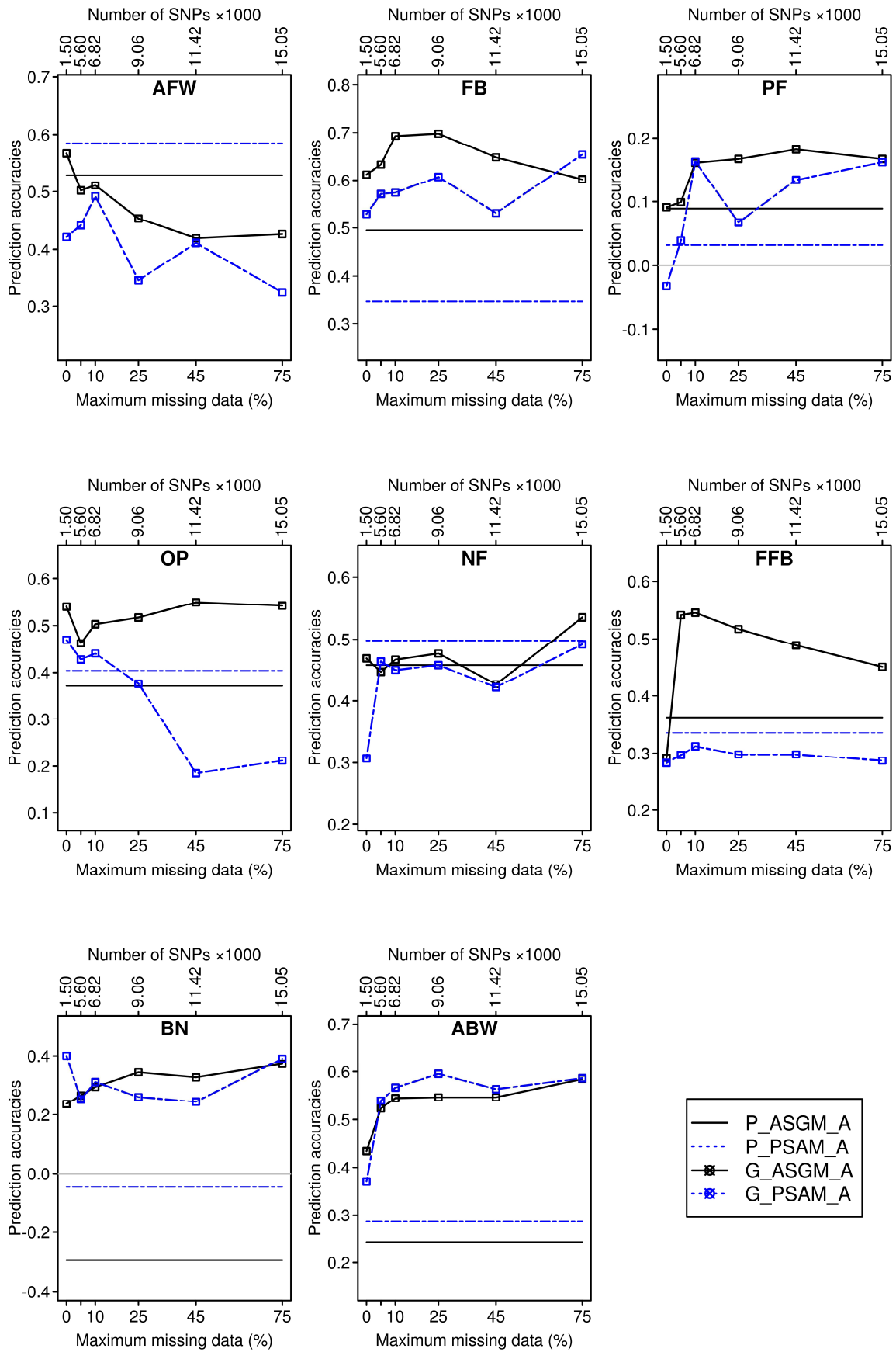
**Fig. 1.** Imputation and phasing scheme for the production of the SNP datasets used for genomic predictions with the two models PSAM (population-specific effects of SNP alleles model) and ASGM (across-population SNP genotype model). pA, pB, AxB: Deli parents, La Mé parents and DelixLa Mé hybrid ortets, <sub>(i)</sub> 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.

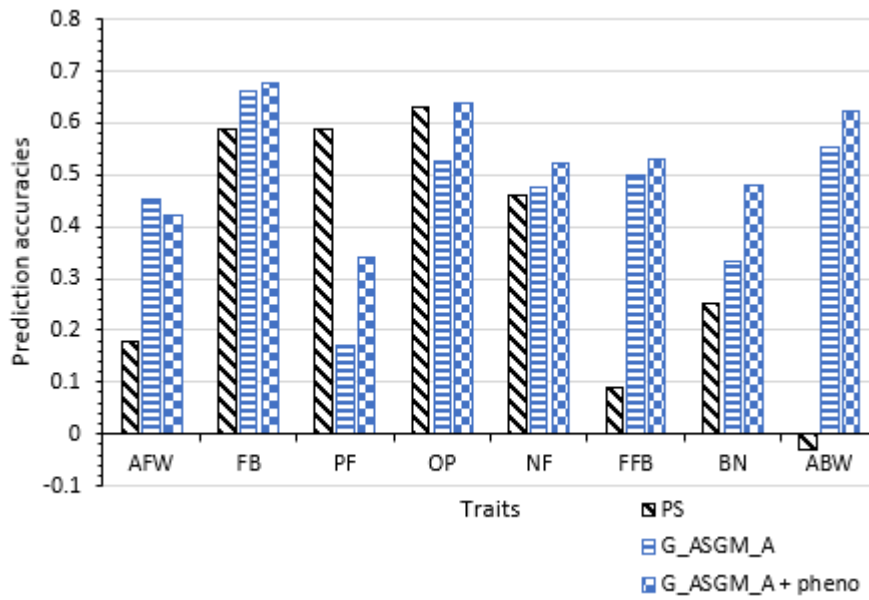
843



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

846

847



848

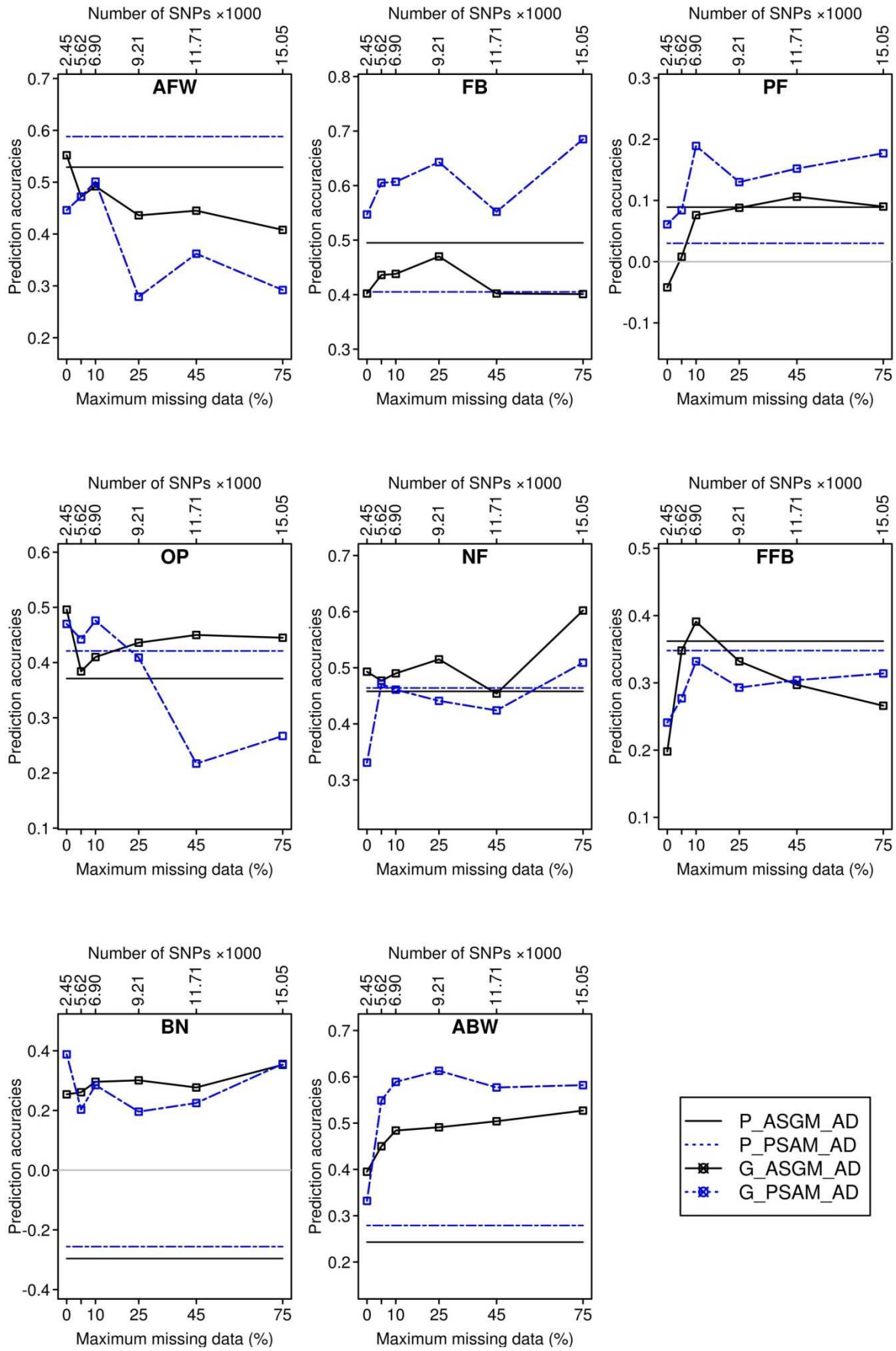
849 **Fig. 4.** Prediction accuracies of phenotypic selection (PS) and of the G\_ASGM\_A model without  
850 phenotypic data (G\_ASGM\_A) and with phenotypic data (G\_ASGM\_A+pheno) of ortets, on average  
851 over the best SNP datasets, and according to trait.

852

853







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