

THE UNIVERSITY OF YAOUNDE I
UNIVERSITE DE YAOUNDE I



FACULTY OF SCIENCE
FACULTE DES SCIENCES

CENTRE FOR RESEARCH AND TRAINING IN GRADUATE STUDIES IN LIFE SCIENCE,
HEALTH & ENVIRONMENTAL SCIENCES
*CENTRE DE RECHERCHE ET DE FORMATION DOCTORALE EN SCIENCES DE LA VIE,
SANTÉ ET ENVIRONNEMENT*

DEPARTMENT OF PLANT BIOLOGY

DÉPARTEMENT DE BIOLOGIE ET PHYSIOLOGIE VÉGÉTALES

**Contribution of genomic selection to improve palm oil yield in
Elaeis guineensis Jacq.**

Thesis submitted in partial fulfilment of requirements for award of a Doctor of Philosophy
Degree (PhD) in Plant Biology
Option: Plant Biotechnologies

By
NYOUMA Achille
MSc in Plant Biology
Registration Number: 10S0201

Supervised by:

CROS David
Researcher, CIRAD

BELL Joseph Martin
Professor



Year 2021



DEPARTEMENT DE BIOLOGIE ET PHYSIOLOGIE VEGETALES
DEPARTMENT OF PLANT BIOLOGY

ATTESTATION DE CORRECTION

Nous soussignés, membres du jury de soutenance de la thèse de Doctorat/PhD en Biologie des Organismes Végétaux option Biotechnologies Végétales de **Monsieur NYOUMA Achille**, Matricule **10S0201**, soutenue publiquement le mercredi 03 Novembre 2021 sur le sujet : « **Contribution of genomic selection to improve palm oil yield in *Elaeis guineensis* Jacq.** » attestons que les corrections conformément aux remarques et recommandations du jury lors de la soutenance de ladite thèse de Doctorat/PhD ont été effectuées par le candidat.

En foi de quoi la présente attestation lui est délivrée afin de servir et valoir ce que de droit./-

Rapporteurs

BELL Joseph Martin
Professeur

CROS David
Ph.D.

Membres

NGONKEU M. Eddy Léonard
Maître de Conférences

NGANDO EBONGUE Georges F.
Directeur de Recherches

KOUAM Éric B.
Maître de Conférences

Président

AMBANG Zachée
Professeur

PROTOCOL LIST

UNIVERSITÉ DE YAOUNDÉ I Faculté des Sciences Division de la Programmation et du Suivi des Activités Académiques		THE UNIVERSITY OF YAOUNDE I Faculty of Science Division of Programming and Follow-up of Academic Affairs
LISTE DES ENSEIGNANTS PERMANENTS		LIST OF PERMANENT TEACHING STAFF

ANNÉE ACADEMIQUE 2019/2020

(Par Département et par Grade)

DATE D'ACTUALISATION 12 Juin 2020

ADMINISTRATION

DOYEN : TCHOUANKEU Jean- Claude, *Maitre de Conférences*

VICE-DOYEN / DPSAA : ATCHADE Alex de Théodore, *Maitre de Conférences*

VICE-DOYEN / DSSE : AJEAGAH Gideon AGHAINDUM, *Professeur*

VICE-DOYEN / DRC : ABOSSOLO Monique, *Maitre de Conférences*

Chef Division Administrative et Financière : NDOYE FOE Marie C. F., *Maitre de Conférences*

Chef Division des Affaires Académiques, de la Scolarité et de la Recherche DAASR : MBAZE MEVA'A Luc Léonard, *Professeur*

1- DÉPARTEMENT DE BIOCHIMIE (BC) (38)			
N°	NOMS ET PRÉNOMS	GRADE	OBSERVATIONS
1	BIGOGA DIAGA Jude	Professeur	En poste
2	FEKAM BOYOM Fabrice	Professeur	En poste
3	FOKOU Elie	Professeur	En poste
4	KANSCI Germain	Professeur	En poste
5	MBACHAM FON Wilfried	Professeur	En poste
6	MOUNDIPA FEWOU Paul	Professeur	Chef de Département
7	NINTCHOM PENLAP V. épouse BENG	Professeur	En poste
8	OBEN Julius ENYONG	Professeur	En poste

9	ACHU Merci BIH	Maître de Conférences	En poste
10	ATOUGHO Barbara Mma	Maître de Conférences	En poste
11	AZANTSA KINGUE GABIN BORIS	Maître de Conférences	En poste
12	BELINGA née NDOYE FOE M. C. F.	Maître de Conférences	Chef DAF / FS
13	BOUDJEKO Thaddée	Maître de Conférences	En poste
14	DJUIDJE NGOUNOUE Marcelline	Maître de Conférences	En poste
15	EFFA NNOMO Pierre	Maître de Conférences	En poste

16	NANA Louise épouse WAKAM	Maître de Conférences	En poste
17	NGONDI Judith Laure	Maître de Conférences	En poste
18	NGUEFACK Julienne	Maître de Conférences	En poste
19	NJAYOU Frédéric Nico	Maître de Conférences	En poste
20	MOFOR née TEUGWA Clotilde	Maître de Conférences	Inspecteur de Service MINESUP
21	TCHANA KOUATCHOUA Angèle	Maître de Conférences	En poste

22	AKINDEH MBUH NJI	Chargé de Cours	En poste
23	BEBOY EDZENGUELE Sara Nathalie	Chargée de Cours	En poste
24	DAKOLE DABOY Charles	Chargé de Cours	En poste
25	DJUIKWO NKONGA Ruth Viviane	Chargée de Cours	En poste
26	DONGMO LEKAGNE Joseph Blaise	Chargé de Cours	En poste
27	EWANE Cécile Anne	Chargée de Cours	En poste
28	FONKOUA Martin	Chargé de Cours	En poste
29	BEBEE Fadimatou	Chargée de Cours	En poste
30	KOTUE KAPTUE Charles	Chargé de Cours	En poste
31	LUNGA Paul KEILAH	Chargé de Cours	En poste
32	MANANGA Marlyse Joséphine	Chargée de Cours	En poste
33	MBONG ANGIE M. Mary Anne	Chargée de Cours	En poste
34	PECHANGOU NSANGOU Sylvain	Chargé de Cours	En poste
35	Palmer MASUMBE NETONGO	Chargé de Cours	En poste

36	MBOUCHE FANMOE Marceline Joëlle	Assistante	En poste
37	OWONA AYISSI Vincent Brice	Assistant	En poste
38	WILFRIED ANGIE Abia	Assistante	En poste

2- DÉPARTEMENT DE BIOLOGIE ET PHYSIOLOGIE ANIMALES (BPA) (48)			
1	AJEAGAH Gideon AGHAINDUM	Professeur	<i>VICE-DOYEN / DSSE</i>
2	BILONG BILONG Charles-Félix	Professeur	Chef de Département
3	DIMO Théophile	Professeur	En Poste
4	DJIETO LORDON Champlain	Professeur	En Poste
5	ESSOMBA née NTSAMA MBALA	Professeur	<i>Vice Doyen/FMSB/UYI</i>
6	FOMENA Abraham	Professeur	En Poste
7	KAMTCHOING Pierre	Professeur	En poste
8	NJAMEN Dieudonné	Professeur	En poste

9	NJIOKOU Flobert	Professeur	En Poste
10	NOLA Moïse	Professeur	En poste
11	TAN Paul VERNYUY	Professeur	En poste
12	TCHUEM TCHUENTE Louis Albert	Professeur	<i>Inspecteur de service Coord.Progr./MINSANTE</i>
13	ZEBAZE TOGOUET Serge Hubert	Professeur	<i>En poste</i>

14	BILANDA Danielle Claude	Maître de Conférences	En poste
15	DJIOGUE Séfirin	Maître de Conférences	En poste
16	DZEUFUET DJOMENI Paul Désiré	Maître de Conférences	En poste
17	JATSA BOUKENG Hermine épouse MEGAPTCHE	Maître de Conférences	En Poste
18	KEKEUNOU Sévilor	Maître de Conférences	En poste
19	MEGNEKOU Rosette	Maître de Conférences	En poste
20	MONY Ruth épouse NTONE	Maître de Conférences	En Poste
21	NGUEGUIM TSOFAK Florence	Maître de Conférences	En poste
22	TOMBI Jeannette	Maître de Conférences	En poste

23	ALENE Désirée Chantal	Chargée de Cours	En poste
26	ATSAMO Albert Donatien	Chargé de Cours	En poste
27	BELLET EDIMO Oscar Roger	Chargé de Cours	En poste
28	DONFACK Mireille	Chargée de Cours	En poste
29	ETEME ENAMA Serge	Chargé de Cours	En poste
30	GOUNOUE KAMKUMO Raceline	Chargée de Cours	En poste
31	KANDEDA KAVAYE Antoine	Chargé de Cours	En poste
32	LEKEUFACK FOLEFACK Guy B.	Chargé de Cours	En poste
33	MAHOB Raymond Joseph	Chargé de Cours	En poste
34	MBENOUN MASSE Paul Serge	Chargé de Cours	En poste
35	MOUNGANG Luciane Marlyse	Chargée de Cours	En poste
36	MVEYO NDANKEU Yves Patrick	Chargé de Cours	En poste
37	NGOUATEU KENFACK Omer Bébé	Chargé de Cours	En poste
38	NGUEMBOK	Chargé de Cours	En poste
39	NJUA Clarisse Yafi	Chargée de Cours	Chef Div. UBA
40	NOAH EWOTI Olive Vivien	Chargée de Cours	En poste
41	TADU Zephyrin	Chargé de Cours	En poste
42	TAMSA ARFAO Antoine	Chargé de Cours	En poste
43	YEDE	Chargé de Cours	En poste

44	BASSOCK BAYIHA Etienne Didier	Assistant	En poste
45	ESSAMA MBIDA Désirée Sandrine	Assistante	En poste
46	KOGA MANG DOBARA	Assistant	En poste
47	LEME BANOCK Lucie	Assistante	En poste
48	YOUNOUSSA LAME	Assistant	En poste

3- DÉPARTEMENT DE BIOLOGIE ET PHYSIOLOGIE VÉGÉTALES (BPV) (33)			
1	AMBANG Zachée	Professeur	Chef Division/UYII
2	BELL Joseph Martin	Professeur	En poste
3	DJOCGOUE Pierre François	Professeur	En poste
4	MOSSEBO Dominique Claude	Professeur	En poste
5	YOUMBI Emmanuel	Professeur	Chef de Département
6	ZAPFACK Louis	Professeur	En poste

7	ANGONI Hyacinthe	Maître de Conférences	En poste
8	BIYE Elvire Hortense	Maître de Conférences	En poste
9	KENGNE NOUMSI Ives Magloire	Maître de Conférences	En poste
10	MALA Armand William	Maître de Conférences	En poste
11	MBARGA BINDZI Marie Alain	Maître de Conférences	CT/ MINESUP
12	MBOLO Marie	Maître de Conférences	En poste
13	NDONGO BEKOLO	Maître de Conférences	<i>CE / MINRESI</i>
14	NGODO MELINGUI Jean Baptiste	Maître de Conférences	En poste
15	NGONKEU MAGAPTCHE Eddy L.	Maître de Conférences	En poste
16	TSOATA Esaïe	Maître de Conférences	En poste
17	TONFACK Libert Brice	Maître de Conférences	En poste

18	DJEUANI Astride Carole	Chargé de Cours	En poste
19	GOMANDJE Christelle	Chargée de Cours	En poste
20	MAFFO MAFFO Nicole Liliane	Chargé de Cours	En poste
21	MAHBOU SOMO TOUKAM. Gabriel	Chargé de Cours	En poste
22	NGALLE Hermine BILLE	Chargée de Cours	En poste
23	NGOUO Lucas Vincent	Chargé de Cours	En poste
24	NNANGA MEBENGA Ruth Laure	Chargé de Cours	En poste
25	NOUKEU KOUAKAM Armelle	Chargé de Cours	En poste
26	ONANA JEAN MICHEL	Chargé de Cours	En poste

27	GODSWILL NTSOMBAH NTSEFONG	Assistant	En poste
28	KABELONG BANAHO Louis-Paul- Roger	Assistant	En poste
29	KONO Léon Dieudonné	Assistant	En poste
30	LIBALAH Moses BAKONCK	Assistant	En poste
31	LIKENG-LI-NGUE Benoit C	Assistant	En poste
32	TAEDOUNG Evariste Hermann	Assistant	En poste
33	TEMEGNE NONO Carine	Assistant	En poste

4- DÉPARTEMENT DE CHIMIE INORGANIQUE (CI) (34)			
1	AGWARA ONDOH Moïse	Professeur	<i>Chef de Département</i>
2	ELIMBI Antoine	Professeur	En poste
3	Florence UFI CHINJE épouse MELO	Professeur	<i>Recteur Univ.Ngaoundere</i>
4	GHOGOMU Paul MINGO	Professeur	<i>Ministre Chargé de Miss.PR</i>
5	NANSEU Njiki Charles Péguy	Professeur	En poste
6	NDIFON Peter TEKE	Professeur	<i>CT MINRESI</i>
7	NGOMO Horace MANGA	Professeur	<i>Vice Chancellor/UB</i>
8	NDIKONTAR Maurice KOR	Professeur	<i>Vice-Doyen Univ. Bamenda</i>
9	NENWA Justin	Professeur	En poste
10	NGAMENI Emmanuel	Professeur	<i>DOYEN FS UD's</i>

11	BABALE née DJAM DOUDOU	Maître de Conférences	<i>Chargée Mission P.R.</i>
12	DJOUFAC WOUMFO Emmanuel	Maître de Conférences	En poste
13	EMADACK Alphonse	Maître de Conférences	En poste
14	KAMGANG YOUNBI Georges	Maître de Conférences	En poste
15	KEMMEGNE MBOUGUEM Jean C.	Maître de Conférences	En poste
16	KONG SAKEO	Maître de Conférences	En poste
17	NDI NSAMI Julius	Maître de Conférences	En poste
18	NJIOMOU C. épouse DJANGANG	Maître de Conférences	En poste
19	NJOYA Dayirou	Maître de Conférences	En poste

20	ACAYANKA Elie	Chargé de Cours	En poste
21	BELIBI BELIBI Placide Désiré	Chargé de Cours	CS/ ENS Bertoua
22	CHEUMANI YONA Arnaud M.	Chargé de Cours	En poste
23	KENNE DEDZO GUSTAVE	Chargé de Cours	En poste
24	KOUOTOU DAOUDA	Chargé de Cours	En poste
25	MAKON Thomas Beauregard	Chargé de Cours	En poste
26	MBEY Jean Aime	Chargé de Cours	En poste
27	NCHIMI NONO KATIA	Chargé de Cours	En poste
28	NEBA nee NDOSIRI Bridget NDOYE	Chargée de Cours	CT/ MINFEM
29	NYAMEN Linda Dyorisse	Chargée de Cours	En poste
30	PABOUDAM GBAMBIE A.	Chargée de Cours	En poste
31	TCHAKOUTE KOUAMO Hervé	Chargé de Cours	En poste
32	NJANKWA NJABONG N. Eric	Assistant	En poste
33	PATOUOSSA ISSOFA	Assistant	En poste
34	SIEWE Jean Mermoz	Assistant	En Poste

5- DÉPARTEMENT DE CHIMIE ORGANIQUE (CO) (35)

1	DONGO Etienne	Professeur	Vice-Doyen
2	GHOGOMU TIH Robert Ralph	Professeur	Dir. IBAF/UDA
3	NGOUELA Silvère Augustin	Professeur	Chef de Département UDS
4	NKENGFACK Augustin Ephrem	Professeur	Chef de Département
5	NYASSE Barthélemy	Professeur	En poste
6	PEGNYEMB Dieudonné Emmanuel	Professeur	<i>Directeur/ MINESUP</i>
7	WANDJI Jean	Professeur	En poste

8	Alex de Théodore ATCHADE	Maître de Conférences	Vice-Doyen / DPSAA
9	EYONG Kenneth OBEN	Maître de Conférences	En poste
10	FOLEFOC Gabriel NGOSONG	Maître de Conférences	En poste
11	FOTSO WABO Ghislain	Maître de Conférences	En poste
12	KEUMEDJIO Félix	Maître de Conférences	En poste
13	KEUMOGNE Marguerite	Maître de Conférences	En poste
14	KOUAM Jacques	Maître de Conférences	En poste
15	MBAZOA née DJAMA Céline	Maître de Conférences	En poste
16	MKOUNGA Pierre	Maître de Conférences	En poste

17	NOTE LOUGBOT Olivier Placide	Maître de Conférences	Chef Service/MINESUP
18	NGO MBING Joséphine	Maître de Conférences	Sous/Direct. MINERESI
19	NGONO BIKOBO Dominique Serge	Maître de Conférences	En poste
20	NOUNGOUE TCHAMO Diderot	Maître de Conférences	En poste
21	TABOPDA KUATE Turibio	Maître de Conférences	En poste
22	TCHOUANKEU Jean-Claude	Maître de Conférences	<i>Doyen /FS/ UYI</i>
23	TIH née NGO BILONG E. Anastasie	Maître de Conférences	En poste
24	YANKEP Emmanuel	Maître de Conférences	En poste

25	AMBASSA Pantaléon	Chargé de Cours	En poste
26	KAMTO Eutrophe Le Doux	Chargé de Cours	En poste
27	MVOT AKAK CARINE	Chargé de Cours	En poste
28	NGNINTEDO Dominique	Chargé de Cours	En poste
29	NGOMO Orléans	Chargée de Cours	En poste
30	OUAHOUE WACHE Blandine M.	Chargée de Cours	En poste
31	SIELINOUE TEDJON Valérie	Chargé de Cours	En poste
32	TAGATSING FOTSING Maurice	Chargé de Cours	En poste
33	ZONDENDEGOUMBA Ernestine	Chargée de Cours	En poste

34	MESSI Angélique Nicolas	Assistant	En poste
35	TSEMEUGNE Joseph	Assistant	En poste

6- DÉPARTEMENT D'INFORMATIQUE (IN) (27)

1	ATSA ETOUNDI Roger	Professeur	<i>Chef Div. MINESUP</i>
2	FOUDA NDJODO Marcel Laurent	Professeur	<i>Chef Dpt ENS/Chef IGA. MINESUP</i>

3	NDOUNDAM René	Maître de Conférences	En poste
---	---------------	-----------------------	----------

4	AMINOUE Halidou	Chargé de Cours	<i>Chef de Département</i>
5	DJAM Xaviera YOUH - KIMBI	Chargé de Cours	En Poste
6	EBELE Serge Alain	Chargé de Cours	En poste
7	KOUOKAM KOUOKAM E. A.	Chargé de Cours	En poste
8	MELATAGIA YONTA Paulin	Chargé de Cours	En poste
9	MOTO MPONG Serge Alain	Chargé de Cours	En poste
10	TAPAMO Hyppolite	Chargé de Cours	En poste
11	ABESOLO ALO'O Gislain	Chargé de Cours	En poste

12	MONTHE DJIADEU Valery M.	Chargé de Cours	En poste
13	OLLE OLLE Daniel Claude Delort	Chargé de Cours	C/D Enset. Ebolowa
14	TINDO Gilbert	Chargé de Cours	En poste
15	TSOPZE Norbert	Chargé de Cours	En poste
16	WAKU KOUAMOU Jules	Chargé de Cours	En poste

17	BAYEM Jacques Narcisse	Assistant	En poste
18	DOMGA KOMGUEM Rodrigue	Assistant	En poste
19	EKODECK Stéphane Gaël Raymond	Assistant	En poste
20	HAMZA Adamou	Assistant	En poste
21	JIOMEKONG AZANZI Fidel	Assistant	En poste
22	MAKEMBE. S. Oswald	Assistant	En poste
23	MESSI NGUELE Thomas	Assistant	En poste
24	MEYEMDOU Nadège Sylvianne	Assistante	En poste
25	NKONDOCK. MI. BAHANACK.N.	Assistant	En poste

7- DÉPARTEMENT DE MATHÉMATIQUES (MA) (30)

1	EMVUDU WONO Yves S.	Professeur	<i>Inspecteur MINESUP</i>
---	---------------------	------------	-------------------------------

2	AYISSI Raoult Domingo	Maître de Conférences	Chef de Département
3	NKUIMI JUGNIA Célestin	Maître de Conférences	En poste
4	NOUNDJEU Pierre	Maître de Conférences	<i>Chef service des programmes & Diplômes</i>
5	MBEHOU Mohamed	Maître de Conférences	En poste
6	TCHAPNDA NJABO Sophonie B.	Maître de Conférences	Directeur/AIMS Rwanda

7	AGHOUKENG JIOFACK Jean Gérard	Chargé de Cours	Chef Cellule MINPLAMAT
8	CHENDJOU Gilbert	Chargé de Cours	En poste
9	DJIADEU NGAHA Michel	Chargé de Cours	En poste
10	DOUANLA YONTA Herman	Chargé de Cours	En poste
11	FOMEKONG Christophe	Chargé de Cours	En poste
12	KIANPI Maurice	Chargé de Cours	En poste
13	KIKI Maxime Armand	Chargé de Cours	En poste
14	MBAKOP Guy Merlin	Chargé de Cours	En poste
15	MBANG Joseph	Chargé de Cours	En poste
16	MBELE BIDIMA Martin Ledoux	Chargé de Cours	En poste
17	MENGUE MENGUE David Joe	Chargé de Cours	En poste
18	NGUEFACK Bernard	Chargé de Cours	En poste

19	NIMPA PEFOUKEU Romain	Chargée de Cours	En poste
20	POLA DOUNDOU Emmanuel	Chargé de Cours	En poste
21	TAKAM SOH Patrice	Chargé de Cours	En poste
22	TCHANGANG Roger Duclos	Chargé de Cours	En poste
23	TCHOUNDJA Edgar Landry	Chargé de Cours	En poste
24	TETSADJIO TCHILEPECK M. E.	Chargée de Cours	En poste
25	TIAYA TSAGUE N. Anne-Marie	Chargée de Cours	En poste
26	MBIAKOP Hilaire George	Assistant	En poste
27	BITYE MVONDO Esther Claudine	Assistante	En poste
28	MBATAKOU Salomon Joseph	Assistant	En poste
29	MEFENZA NOUNTU Thiery	Assistant	En poste
30	TCHEUTIA Daniel Duviol	Assistant	En poste

8- DÉPARTEMENT DE MICROBIOLOGIE (MIB) (18)

1	ESSIA NGANG Jean Justin	Professeur	<i>Chef de Département</i>
---	-------------------------	------------	----------------------------

2	BOYOMO ONANA	Maître de Conférences	En poste
3	NWAGA Dieudonné M.	Maître de Conférences	En poste
4	NYEGUE Maximilienne Ascension	Maître de Conférences	En poste
5	RIWOM Sara Honorine	Maître de Conférences	En poste
6	SADO KAMDEM Sylvain Leroy	Maître de Conférences	En poste

7	ASSAM ASSAM Jean Paul	Chargé de Cours	En poste
8	BODA Maurice	Chargé de Cours	En poste
9	BOUGNOM Blaise Pascal	Chargé de Cours	En poste
10	ESSONO OBOUGOU Germain G.	Chargé de Cours	En poste
11	NJIKI BIKOÏ Jacky	Chargée de Cours	En poste
12	TCHIKOUA Roger	Chargé de Cours	En poste

13	ESSONO Damien Marie	Assistant	En poste
14	LAMYE Glory MOH	Assistant	En poste
15	MEYIN A EBONG Solange	Assistante	En poste
16	NKOUDOU ZE Nardis	Assistant	En poste
17	SAKE NGANE Carole Stéphanie	Assistante	En poste
18	TOBOLBAÏ Richard	Assistant	En poste

9. DEPARTEMENT DE PYSIQUE(PHY) (40)			
1	BEN- BOLIE Germain Hubert	Professeur	En poste
2	EKOBENA FOU DA Henri Paul	Professeur	<i>Chef Division. UN</i>
3	ESSIMBI ZOBO Bernard	Professeur	En poste
4	KOFANE Timoléon Crépin	Professeur	En poste
5	NANA ENGO Serge Guy	Professeur	En poste
6	NDJAKA Jean Marie Bienvenu	Professeur	Chef de Département
7	NOUAYOU Robert	Professeur	En poste
8	NJANDJOCK NOUCK Philippe	Professeur	<i>Sous Directeur/ MINRESI</i>
9	PEMHA Elkana	Professeur	En poste
10	TABOD Charles TABOD	Professeur	Doyen Univ/Bda
11	TCHAWOUA Clément	Professeur	En poste
12	WOAFO Paul	Professeur	En poste

13	BIYA MOTTO Frédéric	Maître de Conférences	DG/HYDRO Mekin
14	BODO Bertrand	Maître de Conférences	En poste
15	DJUIDJE KENMOE épouse ALOYEM	Maître de Conférences	En poste
16	EYEBE FOU DA Jean sire	Maître de Conférences	En poste
17	FEWO Serge Ibraïd	Maître de Conférences	En poste
18	HONA Jacques	Maître de Conférences	En poste
19	MBANE BIOUELE César	Maître de Conférences	En poste
20	NANA NBENDJO Blaise	Maître de Conférences	En poste
21	NDOP Joseph	Maître de Conférences	En poste
22	SAIDOU	Maître de Conférences	MINERESI
23	SIEWE SIEWE Martin	Maître de Conférences	En poste
24	SIMO Elie	Maître de Conférences	En poste
25	VONDOU Derbetini Appolinaire	Maître de Conférences	En poste
26	WAKATA née BEYA Annie	Maître de Conférences	<i>Sous Directeur/ MINESUP</i>
27	ZEKENG Serge Sylvain	Maître de Conférences	En poste

28	ABDOURAHIMI	Chargé de Cours	En poste
29	EDONGUE HERVAIS	Chargé de Cours	En poste
30	ENYEGUE A NYAM épouse BELINGA	Chargée de Cours	En poste
31	FOUEDJIO David	Chargé de Cours	Chef Cell. MINADER
32	MBINACK Clément	Chargé de Cours	En poste
33	MBONO SAMBA Yves Christian U.	Chargé de Cours	En poste
34	MEL'I Joelle Larissa	Chargée de Cours	En poste
35	MVOGO ALAIN	Chargé de Cours	En poste
36	OBOUNOU Marcel	Chargé de Cours	DA/Univ Inter Etat/Sangmalima
37	WOULACHE Rosalie Laure	Chargée de Cours	En poste

38	AYISSI EYEBE Guy François Valérie	Assistant	En poste
39	CHAMANI Roméo	Assistant	En poste
40	TEYOU NGOUPOU Ariel	Assistant	En poste

10- DÉPARTEMENT DE SCIENCES DE LA TERRE (ST) (43)

1	BITOM Dieudonné	Professeur	<i>Doyen / FASA / UDs</i>
2	FOUATEU Rose épouse YONGUE	Professeur	En poste
3	KAMGANG Pierre	Professeur	En poste
4	NDJIGUI Paul Désiré	Professeur	Chef de Département
5	NDAM NGOUPAYOU Jules-Remy	Professeur	En poste
6	NGOS III Simon	Professeur	DAAC/Uma
7	NKOUMBOU Charles	Professeur	En poste
8	NZENTI Jean-Paul	Professeur	En poste

9	ABOSSOLO née ANGUE Monique	Maître de Conférences	<i>Vice-Doyen / DRC</i>
10	GHOGOMU Richard TANWI	Maître de Conférences	CD/Uma
11	MOUNDI Amidou	Maître de Conférences	<i>CT/ MINIMDT</i>
12	NGUEUTCHOUA Gabriel	Maître de Conférences	CEA/MINRESI
13	NJILAH Isaac KONFOR	Maître de Conférences	En poste
14	ONANA Vincent Laurent	Maître de Conférences	<i>Chef service Maintenance & du Matériel</i>
15	BISSO Dieudonné	Maître de Conférences	<i>Directeur/Projet Barrage Memve'ele</i>
16	EKOMANE Emile	Maître de Conférences	En poste
17	GANNO Sylvestre	Maître de Conférences	En poste
18	NYECK Bruno	Maître de Conférences	En poste

19	TCHOUANKOUE Jean-Pierre	Maître de Conférences	En poste
20	TEMDJIM Robert	Maître de Conférences	En poste
21	YENE ATANGANA Joseph Q.	Maître de Conférences	<i>Chef Div. /MINTP</i>
22	ZO'O ZAME Philémon	Maître de Conférences	<i>DG/ART</i>

23	ANABA ONANA Achille Basile	Chargé de Cours	En poste
24	BEKOA Etienne	Chargé de Cours	En poste
25	ELISE SABABA	Chargé de Cours	En poste
26	ESSONO Jean	Chargé de Cours	En poste
27	EYONG JOHN TAKEM	Chargé de Cours	En poste
28	FUH Calistus Gentry	Chargé de Cours	<i>Sec. D'Etat/MINMIDT</i>
29	LAMILEN BILLA Daniel	Chargé de Cours	En poste
30	MBESSE CECILE OLIVE	Chargée de Cours	En poste
31	MBIDA YEM	Chargé de Cours	En poste
32	METANG Victor	Chargé de Cours	En poste
33	MINYEM Dieudonné-Lucien	Chargé de Cours	<i>CD/Uma</i>
34	NGO BELNOUN Rose Noël	Chargée de Cours	En poste
35	NGO BIDJECK Louise Marie	Chargée de Cours	En poste
36	NOMO NEGUE Emmanuel	Chargé de Cours	En poste
37	NTSAMA ATANGANA Jacqueline	Chargée de Cours	En poste
38	TCHAKOUNTE J. épse NOUMBEM	Chargée de Cours	<i>Chef.cell / MINRESI</i>
39	TCHAPTCHET TCHATO De P.	Chargé de Cours	En poste
40	TEHNA Nathanaël	Chargé de Cours	En poste
41	TEMGA Jean Pierre	Chargé de Cours	En poste
42	FEUMBA Roger	Assistant	En poste
43	MBANGA NYOBE Jules	Assistant	En poste

Répartition chiffrée des Enseignants de la Faculté des Sciences de l'Université de Yaoundé I

NOMBRE D'ENSEIGNANTS					
DÉPARTEMENT	Professeurs	Maîtres de Conférences	Chargés de Cours	Assistants	Total
BCH	9 (1)	13 (09)	14 (06)	3 (2)	39 (18)
BPA	13 (1)	09 (06)	19 (05)	05 (2)	46 (14)
BPV	06 (0)	11 (02)	9 (06)	07 (01)	33 (9)
CI	10 (1)	9 (02)	12 (02)	03 (0)	34 (5)
CO	7 (0)	17 (04)	09 (03)	02 (0)	35(7)
IN	2 (0)	1 (0)	13 (01)	09 (01)	25 (2)
MAT	1 (0)	5 (0)	19 (01)	05 (02)	30 (3)
MIB	1 (0)	5 (02)	06 (01)	06 (02)	18 (5)
PHY	12 (0)	15 (02)	10 (03)	03 (0)	40 (5)

ST	8 (1)	14 (01)	19 (05)	02 (0)	43(7)
Total	69 (4)	99 (28)	130 (33)	45 (10)	343 (75)
Soit un total de		344 (75) dont :			
-	Professeurs	68 (4)			
-	Maîtres de Conférences	99 (28)			
-	Chargés de Cours	130 (33)			
-	Assistants	46 (10)			
() = Nombre de Femmes		75			

DEDICATION

This work is dedicated to my late parents

Mr. NDÉMÉ ONANA Robert and Mrs. NDÉMÉ ONANA née KÉYI Geneviève

ACKNOWLEDGMENTS

I would like to warmly thank for their support for the accomplishment of this work many persons and institutions. My gratitude is expressed towards:

- Professor BELL Joseph Martin, my Supervisor, for his invaluable scientific support, his listening, availability, and the trust he kindly placed upon me by accepting me as PhD student. Moreover, he is the one who taught me the basics of scientific research. May he find through the current document the expression of my utmost gratitude;
- Doctor CROS David, my main Advisor, for his tireless and daily monitoring of this thesis work, his advice, and for trusting me. Working under his supervision throughout my PhD study has been an uplifting experience. His passion, enthusiasm, and hardworking capacities encouraged me to progress. I hope he will find through this work the expression of my profound gratitude. I would like through him, to thank the French agricultural research and international cooperation organization (CIRAD) for recruiting me as a PhD intern following a national competitive examination entrance. Moreover, I am grateful for the office with internet connection that they kindly put at my disposal throughout my PhD study;
- Professor YOUMBI Emmanuel, the Head of the Department of Plant Biology for his intellectual rigor and his commitment to the training of students;
- Doctor JACOB Florence, Researcher at PalmElit SAS company, for her thorough review of all the scientific papers of this thesis work. Through her, I would like to thank PalmElit SAS company for providing phenotypic and molecular data and for funding my allowances, travels, university tuition fees, hotel and living costs during conferences, workshops, and stays abroad;
- Professor TEWA Jules of the African Centre of Excellence in Information and Communication Technologies (CETIC), for providing me with an office. Through him, I would like to thank the CETIC project for lending me a brand-new last generation computer and for their financial support;
- Doctors MOURNET Pierre, head of the *Grand Plateau Technique Régional de Génotypage* of CIRAD, Montpellier, for welcoming me and thoroughly presenting me the platform;
- Doctor BILLOTTE Nobert, for inviting me at CIRAD Montpellier;
- Doctor NGALLE Hermine BILLE, for her tireless encouragement and her wise advice during the realization of this work;

- Mr. EYA'A NGOMBO Clément, for informing me of the competitive entrance examination launched by CIRAD/PALMELIT to recruit one PhD student in Cameroon;
- the Genetics and Plant Improvement Unit (UGAP) of the Department of Plant Biology of the University of Yaoundé I, for the relaxing warm and cozy atmosphere, the follow-up, and corrections during my various presentations;
 - all the professors of the Department of Plant Biology, for the lectures and training, received throughout my academic path;
 - Doctor LIKENG-LI-NGUE Benoît Constant, my academic senior, for his support, encouragement, and advice for the realization of this work;
 - Mrs. CHIMI Pierre and AKOA Victor, my classmates for their warmth and assistance;
 - Mr. MUNYENGWA Norman, for partly proofreading this manuscript;
 - My brothers and sisters KÉYI Raïssa, KÉNEMBÉNI Dominique, AMANG André, MBOUSSI Claude-François, ONANA NDÉMÉ Ismaël, BOGONDO Julien and BASSANG'NA Jean-Pierre, for their warmth and encouragement;
 - all those who directly or indirectly contributed to the realization of this work and whose names, unfortunately, do not appear here, may they find here the testimony of my esteem and my deep gratitude;
 - all the members of the jury, for their interest in this research by agreeing to examine and improve the quality of this work.

TABLE OF CONTENTS

PROTOCOL LIST	i
DEDICATION	xiv
ACKNOWLEDGMENTS	xv
TABLE OF CONTENTS	xvii
LIST OF FIGURES	xx
LIST OF TABLES	xxi
LIST OF ABBREVIATIONS	xxii
LIST OF APPENDICES	xxiv
ABSTRACT	xxv
RESUMÉ.....	xxvii
CHAPTER I. GENERALITIES	1
I.1. Introduction	1
I.2. Literature review	5
I.2.1. Generalities on oil palm	5
I.2.1.1. Classification and origin of the oil palms	5
I.2.1.2. Taxonomy and botanical description of <i>E. guineensis</i>	7
I.2.1.3. Oil palm ecology	8
I.2.1.5. Oil palm and environment.....	8
I.2.1.4. Production and economic importance	9
I.2.2. Concepts of quantitative genetics	10
I.2.2.1. Quantitative traits	10
I.2.2.2. Main properties of quantitative traits	11
I.2.2.3. Phenotypic value	11
I.2.2.4. Phenotypic variance	17
I.2.2.5. Resemblance between relatives.....	20
I.2.3. Overview of oil palm genetics and breeding strategies	26
I.2.3.1. Oil palm breeding goals and objectives	26
I.2.3.2. Genetic determinism and fruit forms	27
I.2.3.3. Fruit types.....	28

I.2.3.4. Mantled fruit type.....	29
I.2.3.5. Reproduction system.....	30
I.2.3.6. Genetic resources for oil palm breeding	31
I.2.3.7. Mass selection	32
I.2.3.8. Current breeding schemes	33
I.2.4. Genomic selection.....	41
I.2.4.1. Principle	43
I.2.4.2. Molecular data.....	44
I.2.4.3. Training and application populations.....	45
I.2.4.4. Models and statistical methods for genomic predictions	46
I.2.4.5. Information captured by markers	48
I.2.5. Genetic progress.....	48
CHAPTER II. MATERIAL AND METHODS	51
II.1. Material.....	51
II.1.1. Study sites and experimental designs	51
II.1.2. Plant material.....	52
II.2. Methods	54
II.2.1. Evaluation of the efficiency of genomic selection for clonal selection.....	54
II.2.1.1. Phenotyping	54
II.2.1.2. Genotyping.....	54
II.2.1.3. Imputation of missing SNP data and phasing	55
II.2.1.4. Definition of SNP datasets.....	56
II.2.1.5. Prediction models and computation of genetic values of unobserved clones.....	56
II.2.1.6. Prediction accuracies	61
II.2.1.7. Determination of the reference clonal values predicted by the model.....	61
II.2.1.8. Accuracy of phenotypic selection before clonal trials	62
II.2.2. Effect of the genotyping strategy to optimize prediction accuracy	62
II.2.2.1. Phenotyping	62
II.2.2.2. Generation of SNP molecular data	62
II.2.2.3. Imputation of missing SNP genotypes and phasing	63
II.2.2.4. Models for prediction of hybrid performances	63
II.2.2.5. Reference genetic values of hybrid crosses	66
II.2.2.6. Prediction accuracies and model comparison of models	66

CHAPTER III. RESULTS AND DISCUSSION	69
III.1. Results	69
III.1.1. Efficiency of genomic selection for clonal selection	69
III.1.1.1. Distribution of frequencies of minor and alternate alleles across population	69
III.1.1.2. Effect of GS prediction model and SNP dataset on prediction accuracy	72
III.1.1.3. Comparison of prediction accuracies of PS and GS.....	84
III.1.2. Effect of the genotyping strategy to optimize prediction accuracy.....	86
III.1.2.1. Effect on prediction accuracy of using the genotyping strategy for the training population.....	86
III.1.2.2. Effect on prediction accuracy of the method used to model marker effects .	91
III.1.2.3. Comparison of GS models and control pedigree-based models.....	91
III.2. Discussion	92
III.2.1. Efficiency of genomic selection for clonal selection	93
III.2.1.1. Improving the genetic progress of clonal breeding with GS.....	93
III.2.1.2. Effects of prediction model and SNP dataset on prediction accuracies	94
III.2.1.3. Genotyped individuals for training.....	97
III.2.1.4. Prediction of dominance effects	98
III.2.2. Effect of the genotyping strategy to optimize prediction accuracy.....	98
III.2.2.1. Using genomic data of hybrid individuals to train the GS model	98
III.2.2.2. Effect of modelling of marker on prediction accuracy.....	100
CHAPTER IV. CONCLUSION, PERSPECTIVES AND RECOMMENDATIONS	102
IV.1. Conclusion	102
IV.2. Perspectives.....	103
IV.3. Recommendations	103
REFERENCES.....	104
APPENDICES.....	116

LIST OF FIGURES

Fig. 1. Oil palm tree	6
Fig. 2. Comparison of oil palm pollen..	6
Fig. 3. Distribution of oil palm production worldwide	10
Fig. 4. Genotypic value based on one locus-genotype, with randomly assigned alleles	12
Fig. 5. Allele substitution and genotypic values of the resulting genotypes.....	14
Fig. 6. Decomposition of phenotypes.....	17
Fig. 7. Transmission of identical by descent segment of chromosome in two offspring.....	22
Fig. 8. Inheritance of two identical segments from a common ancestor in an inbred individual	24
Fig. 9. Oil palm fruit forms	28
Fig. 10. Oil palm fruit types	29
Fig. 11. Transversal and longitudinal section of oil palm fruit	30
Fig. 12. Scheme of one cycle of modified reciprocal recurrent selection applied to oil palm. 34	
Fig. 13. Possible scheme of genomic modified reciprocal recurrent selection	42
Fig. 14. Diagram of genomic selection	43
Fig. 15. Map of the study area.....	51
Fig. 16. Location plan of the 28 trials of Aek Loba Timur.....	52
Fig. 17. Imputation and phasing scheme.....	56
Fig. 18. Heat map of additive realized relationships matrices of the 123 parents A of the training set.	67
Fig. 19. Heat map of additive realized relationships matrices of the 121 parents B of the training set.	68
Fig. 20. Distribution of minor allele frequency.....	69
Fig. 21. Correlation of minor allele frequency.....	71
Fig. 22. Prediction accuracies of bunch production traits according to SNP datasets and prediction models.	73
Fig. 23. Prediction accuracies according to traits, SNP datasets and prediction models.	76
Fig. 24. Prediction accuracies of bunch production traits according to SNP datasets and prediction models	78
Fig. 25. Prediction accuracies according to traits, SNP datasets and prediction models.	80
Fig. 26. Prediction accuracies on average over the best SNP datasets and according to trait..	85
Fig. 27. Prediction accuracies of bunch quality traits according to prediction models.....	87
Fig. 28. Prediction accuracies of bunch production traits according to prediction models	89
Fig. 29. Average prediction accuracies of prediction models across traits	90

LIST OF TABLES

Table I. Deduction of the population genotypic mean from the relative allele frequencies and genotypic value	13
Table II. Kinship and fraternity coefficients according to their family relationship.....	21
Table III. Origin of heterosis in oil palm for bunch yield.	36
Table IV. Characteristics of the datasets used for training and validation for clones.....	53
Table V. Composition of the datasets used for training and validation for hybrids.	54
Table VI. Characteristics of the SNP datasets defined based on a threshold in terms of maximum percentage of missing data per individual.....	57
Table VII. Composition of training and validation sets.	62
Table VIII. Characteristics of genotyped hybrid individuals of the training set.....	63
Table IX. Mean prediction accuracies according to trait and prediction model.	81
Table X. Pairwise comparison of prediction accuracies among genomic selection and pedigree-based models, according to SNP dataset and trait.....	82
Table XI. Pairwise comparison of prediction accuracies among genomic selection and pedigree-based models, according to SNP dataset and trait.....	83
Table XII. Intensity and accuracy of phenotypic selection before clonal trials according to trait.	86
Table XIII. Maximum prediction accuracies of traits.	92

LIST OF ABBREVIATIONS

ABW: Average Bunch Weight

AFW: Average Fruit Weight

AVROS: *Algemeene Vereniging van Rubberplantera ter Oostkust van Sumatra*

BLUP: Best Linear Unbiased Predictions

BN: Bunch Number

BPRO: Breeding Populations of Restricted Origin

CNRA: Centre National de Recherche Agronomique de Côte d'Ivoire

CPO: Crude Palm Oil

CRAPP: Centre de Recherches Agricoles Plantes Pérennes du Bénin

FB: Fruit to Bunch ratio

FFB: Fresh Fruit Bunches

FIPS: Family and Individual Palm Selection

GBLUP: Genomic Best Linear Unbiased Predictions

GCA: General Combining Ability

GEBV: Genomic Estimated Breeding Values

GEGV: Genomic Estimated Genetic Values

GS: Genomic Selection

INRAB: Institut National des Recherches Agricoles du Bénin

IOPRI: Indonesian Oil Palm Research Institute

IRAD: Institute of Agricultural Research for Development

MAS: Marker-Assisted Selection

MRRS: Modified Reciprocal Recurrent Selection

MRS: Modified Recurrent Selection

NF: Number of Fruits per bunch

OP: Oil to Pulp ratio

PF: Pulp to Fruit ratio

QTL: Quantitative Trait Locus

RRS: Reciprocal Recurrent Selection

SCA: Specific Combining Ability

SOCFINDO: Société Financière des Caoutchoucs d'Indonésie

LIST OF APPENDICES

Appendix 1. Objectives and corresponding published papers.	116
Appendix 2. Logical framework of objective 1	116
Appendix 3. Logical framework of objective 2	118
Appendix 4. Generation of SNP molecular data	119
Appendix 5. Steps of genotyping-by-sequencing (GBS) in plants	121
Appendix 6. Genetic map of oil palm.	122
appendix 7. Published papers.	124

ABSTRACT

Genomic selection (GS) is expected to increase the annual genetic progress and lead palm oil production up to the growing world demand. Genetic improvement for hybrid performances has a major role to play to meet this demand while minimizing environmental impacts. A modified reciprocal recurrent scheme is used to select the most performing hybrids commercialized as hybrid cultivars or used for the most performing individuals as hybrid ortets in clonal selection. The current study empirically evaluated the interest of using genomic data from A × B hybrid individuals for the genomic approach applied to oil palm (*Elaeis guineensis* Jacq.).

The efficiency of GS for clonal selection was first evaluated using a training set comprising 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, across-population SNP genotype models (ASGM) and population-specific effects of SNP alleles models (PSAM), respectively ignoring considering the parental origin of alleles, were assessed. Secondly, we investigated the effect of two strategies to optimize the GS accuracy in oil palm hybrid: genotyping strategy for the training population, i.e., genotyping only the hybrid parents or also a sample of hybrid individuals, and modeling of markers ASGM and PSAM. For that purpose, genomic data of both parents and hybrid individuals were used for calibration and predictions were done using ASGM and PSAM. The training set was constructed with around 350 hybrid crosses, including around 15,000 to 23,000 individuals phenotyped, depending on trait. Validation was realized in an independent set of 213 hybrid crosses. GBS was applied on the parents of the training and validation sets and on around 400 training hybrid individuals, yielding 21,458 SNPs.

The results showed prediction accuracies ranging from 0.08 to 0.70 for ortet candidates without data records, depending on trait, SNP dataset and modeling. ASGM with a mean prediction of 0.45 was better (on average slightly more accurate, less sensitive to SNP dataset and simpler) than PSAM with a mean prediction accuracy of 0.43, although PSAM appeared interesting for a few traits. With ASGM, the number of SNPs had to reach 7,000, while the percentage of missing data per SNP was of secondary importance, and GS prediction accuracies were higher than those of PS for most of the traits.

Prediction accuracies ranged from 0.15–0.89 for hybrid crosses depending on trait, model and genotyping strategy. Prediction accuracies increased on average by 5% when

training was done with genomic data of hybrid individuals and parents compared with only parental genomic data. Prediction accuracies increased on average by 3% with ASGM compared to PSAM. In our dataset, the mean prediction accuracy over traits of the best GS approach, i.e., ASGM with hybrid individuals' genotypes, reached 0.53.

Ultimately, this work makes possible two practical applications of GS, that will increase genetic progress by improving ortet preselection before clonal trials: preselection at the mature stage on all yield components jointly using ortet genotypes and phenotypes, and genomic preselection on more yield components than PS, among a large population of the best possible crosses at nursery stage. In addition, this work revealed that genomic data of the training hybrid individuals and GBLUP are useful to increase prediction accuracy; with ASGM the recommended modeling approach for that purpose. Further studies should investigate the factors controlling the relative performance of ASGM and PSAM approaches in oil palm, and focus on the optimal number of hybrid individuals to genotype to maximize the selection response per unit cost.

Keywords: *Elaeis guineensis* Jacq., genomic selection, clonal selection, genotyping-by-sequencing, prediction accuracy.

RESUMÉ

La sélection génomique (SG) peut augmenter le progrès génétique annuel et la production en huile de palme afin de satisfaire la demande mondiale croissante. L'amélioration génétique des performances des hybrides a un rôle majeur à jouer pour répondre à cette demande tout en minimisant les impacts environnementaux. Le schéma de sélection réciproque est utilisé afin de sélectionner les hybrides les plus performants qui sont commercialisés comme cultivars ou alors pour les meilleurs de ces individus, utilisés comme têtes de clone (ortets) dans la sélection clonale. La présente étude a évalué empiriquement l'intérêt de l'utilisation des données génomiques d'individus hybrides A × B pour l'approche génomique appliquée au palmier à huile (*Elaeis guineensis* Jacq.).

D'une part, l'efficacité de la SG pour la sélection clonale a d'abord été évaluée à l'aide d'une population de calibration comprenant près de 300 croisements Deli × La Mé, phénotypés pour huit composantes de rendement en huile de palme et la population de validation comprenant 42 ortets Deli × La Mé. Le génotypage par séquençage (GBS) a révélé 15 054 polymorphismes mono-nucléotidiques (SNP). L'effet des jeux de données SNP (densité et pourcentage de données manquantes) et de deux approches de modélisation de la SG ont été évalués : les modèles de génotypes SNPs à travers la population (ASGM) et les modèles des effets spécifiques aux allèles SNPs de la population (PSAM) ; ignorant et prenant en compte l'origine parentale des allèles respectivement.

D'autre part, l'effet de deux stratégies d'optimisation de la précision de la SG chez les hybrides de palmier à huile a été examiné : stratégie de génotypage pour la population de calibration, c'est-à-dire, génotypage des parents hybrides uniquement ou génotypage également d'un échantillon d'individus hybrides, et modélisation ASGM et PSAM. Les prédictions ont été effectuées à l'aide des modèles ASGM et PSAM qui ont été calibrés en utilisant les données génomiques des parents et des individus hybrides. La population de calibration a été construite avec environ 350 croisements hybrides, soit environ 15 000 à 23 000 individus phénotypés, selon les caractères. La validation a été réalisée sur une population indépendante de 213 croisements hybrides. Le GBS a été appliqué sur les parents des populations de calibration et de validation, et sur environ 400 individus hybrides de la population de calibration, générant ainsi 21 458 SNPs.

Les résultats révèlent des précisions de prédiction allant de 0,08 à 0,70 pour les ortets sans leurs phénotypes, en fonction des caractères, du jeu de données SNP et de l'approche de

modélisation. Le modèle ASGM avec une précision moyenne de 0.45 est meilleur (légèrement plus précis, moins sensible au jeu de données SNP et plus simple) que le modèle PSAM avec une précision moyenne de 0.43, bien que PSAM semble intéressant pour trois caractères. Environ 7 000 SNPs sont nécessaires lorsque le modèle ASGM est utilisé, alors que le pourcentage de données manquantes par SNP est d'importance secondaire, et les précisions de prédiction de la SG sont plus élevées que celles de la sélection phénotypique (SP) pour la plupart des caractères.

Les précisions de prédiction vont de 0,15 à 0,89 pour les croisements hybrides en fonction des caractères, du modèle et de la stratégie de génotypage. Les précisions de prédiction augmentent en moyenne de 5 % lorsque la calibration est effectuée avec des données génomiques d'individus hybrides et de parents par rapport à une calibration effectuée uniquement avec les données génomiques parentales. Les précisions de prédiction augmentent en moyenne de 3 % avec ASGM par rapport à PSAM. La précision de prédiction moyenne sur les caractères de la meilleure approche de SG, c'est-à-dire ASGM avec des génotypes d'individus hybrides, est de 0,53.

En définitive, cette étude permet deux applications pratiques de la SG qui augmenteront le progrès génétique en améliorant la présélection d'ortets avant les essais clonaux : la présélection au stade mature sur toutes les composantes du rendement en utilisant conjointement des génotypes et phénotypes des ortets, et la présélection génomique sur plus de composants de rendement que la SP, parmi une large population des meilleurs croisements possibles au stade pépinière. Par ailleurs, ces travaux révèlent que calibrer les modèles de SG avec un échantillon de données génomiques d'individus hybrides en plus de celles des parents et l'utilisation du modèle ASGM sont d'une grande importance pour augmenter la précision des prédictions. D'autres études sont nécessaires pour examiner les facteurs contrôlant la performance relative des approches ASGM et PSAM chez le palmier à huile, et se focaliser sur le nombre optimal d'individus hybrides à génotyper afin de maximiser la réponse à sélection en fonction du coût.

Mots clés : *Elaeis guineensis* Jacq., sélection génomique, sélection clonale, génotypage par séquençage, précision de prédiction.

CHAPTER I. GENERALITIES

I.1. Introduction

Genomic selection (GS) (Meuwissen *et al.*, 2001) 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) (Goddard & Hayes, 2007). 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 effects. The GS model is calibrated (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 set of 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 (Grattapaglia *et al.*, 2018). 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. Therefore, for a given species, GS allows selecting elite individuals based only on their genomic information, thus, making possible the shortening of the breeding cycle and/or the increase of selection intensity.

Oil palm (*Elaeis guineensis* Jacq.), an allogamous species of the Arecaceae family, is the main oleaginous worldwide through its annual yield of four tons of crude palm oil (CPO) per hectare and a world production above 75 million tons CPO (Anonymous, 2020c). Oil palm production is 36% of the world's vegetable oils on only 0.36% of the world's agricultural lands (Mayes, 2020). Most cultivated oil palms 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 leading to a 30% yield increase explains why they were widely adopted in the 1960s (Corley & Tinker, 2016). The commercial oil palm material is *tenera* (thin-shelled) fruit form, resulting from the cross between the thick-shelled *dura* of group A and the shell-less and usually female sterile *pisifera* of group B. Selection of hybrids is carried out through progeny tests in a

modified reciprocal recurrent selection (MRRS) breeding scheme (Gascon & Berchoux, 1964; Meunier & Gascon, 1972). The best hybrids are primarily selected based on the parental general combining abilities (GCA). While progeny-testing has the advantage of providing high prediction accuracy, it also lengthens the selection cycle by up to ten years. That enabled an annual genetic progress of 1–1.5% so far (Hardon *et al.*, 1987; Soh *et al.*, 2003; Rival & Levang, 2014). Although the annual yield of the oil palm hybrids obtained through the genetic improvement of A×B hybrids increased over the past decades (Rival & Levang, 2014), this remains insufficient to face the expected increase in the demand. Therefore, an additional yield increase is expected. Indeed, the world population is expected to be over nine billion by 2050, and the annual demand for palm oil to be between 120 and 156 million tons (Corley, 2009; Rival & Levang, 2014). Genetic improvement has a major role to play to meet this demand while minimizing environmental impacts. The so far used commercial A×B *tenera* hybrids essentially take advantage of the between-hybrid crosses variability. However, the within-hybrid crosses genetic variability (additive and non-additive) can be exploited in two ways to increase the genetic gain.

Firstly, a supplementary 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 *tenera* individuals (ortets) (Corley & Law, 1997). 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 (Jaligot *et al.*, 2000; Ong-Abdullah *et al.*, 2015; Soh *et al.*, 2017).

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 (Ong-Abdullah *et al.*, 2015), 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. Indeed, Nouy *et al.* (2006) found a broad-sense heritability (H^2) of 0 and 0.1 for bunch number and total bunch production, respectively, thus making the estimation of their genetic values 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, Nouy *et al.* (2006) 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.

Secondly, taking advantage of the within-crosses genetic variability to increase the prediction accuracy can lead to an additional yield increase of sexual crosses for outcrossing species where hybrid parents are heterozygotes (e.g. in oil palm (Nyouma *et al.*, 2019), eucalyptus (Bouvet *et al.*, 2016), robusta coffee (Leroy *et al.*, 1997), etc) depending on the genotyping strategy. When the progeny-tested hybrid parents are in sufficient number to form a training set, two genotyping strategies are possible for the training set: genotyping only the hybrid parents, in order to reduce the genotyping costs, and genotyping also hybrid individuals, or at least a sample. To our knowledge, such a comparison was not made yet.

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 (Wong & Bernardo, 2008), and the possibility of increasing the performance of sexual hybrid crosses by genomic preselection before progeny-tests (Cros *et al.*, 2017). 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 (Nyouma *et al.*, 2019). So far, GS has been successfully used in oil palm (Cros *et al.*, 2015a,b, 2017, 2018; Kwong *et al.*, 2017a) parent selection of hybrid individuals (Cros *et al.*, 2017; Kwong *et al.*, 2017a). A previous empirical study predicted hybrid phenotypes using a thousand hybrid individuals as a training set (Kwong *et al.*, 2017a). Although phenotypes are estimates of the total genetic values, 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, despite the potential benefits that genomic clonal selection have already shown in other perennial crops such as the eucalyptus (Durán *et al.*, 2017) and the rubber tree (Cros *et al.*, 2019). In addition, while genotypes of hybrid individuals take profit of the within-crosses variability, they however, present the major drawback of being expensive given the large number of hybrid individuals to genotype, thus reducing the economic interest of using GS.

Moreover, in a simulation study, Cros *et al.* (2015a) demonstrated that including genomic data of a set of hybrid individuals (1,000) in addition to those of their parents

significantly increase genomic prediction accuracies. Such studies are common in animal breeding (Xiang *et al.*, 2016). However, to our knowledge in plant breeding, no empirical study of that kind has already been performed despite the potential benefits in terms of prediction accuracy and genetic gain that it could provide. To value such type of genomic data, appropriate modeling approaches and imputation and phasing methods will be of great interest.

Given that hybrid cultivars or ortets for clonal selection come from a cross between two oil palm origins, the genomic prediction of their genetic values can be done using two modeling approaches (Ibáñez-Escriche *et al.*, 2009), which are the genomic extensions of the modeling approach developed by Stuber & Cockerham (1966) for interpopulation hybrids. The first approach, the population-specific effects of single nucleotide polymorphism (SNP) alleles model (PSAM, or breed-specific effects of SNP alleles model (BSAM) in the animal breeding literature), 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 (Ibáñez-Escriche *et al.*, 2009). Only a few articles investigated this aspect in animals (Ibáñez-Escriche *et al.*, 2009; Stock *et al.*, 2020). However, to our knowledge, in the context of plant hybrids, these types of models were only compared in simulated maize populations (Technow *et al.*, 2012).

Based on the above, it is legitimate to ask how we could (better) exploit within hybrid crosses variability to improve genetic gain and therefore palm oil yield.

From this overall question, it emerges specific questions such as:

- how can we improve the prediction of the genetic values of A×B hybrid individuals for a better clonal selection in oil palm?
- with regard to recent simulation studies on oil palm, can training using genomic data from parents and hybrid individuals improve the prediction of the genetic values of parents A and B for yield components in oil palm?

The hypotheses resulting from these objectives are:

- training genomic selection models using genomic data from ortets and parents improves the prediction of the genetic value of A×B hybrid individuals for clonal selection in oil palm;

- training genomic selection models using genomic data from parents and hybrid individuals improves the prediction of the genetic values of parents A and B for yield components in oil palm.

The general objective of this study is to evaluate empirically the interest of using genomic data from A × B hybrid individuals for the genomic approach applied to oil palm. The specific objectives are:

- to evaluate the efficiency of genomic selection for clonal selection;
- to investigate the effect of the genotyping strategy to optimize prediction accuracy.

I.2. Literature review

I.2.1. Generalities on oil palm

I.2.1.1. Classification and origin of the oil palms

The genus *Elaeis* comprises two main species whose study is of some interest both economically and genetically: the cultivated African oil palm *E. guineensis* Jacq. and the American oil palm *E. oleifera* (HBK) Cortès. Two other species namely *E. madagascariensis* and *E. odorata* are sometimes evoked in literature but present a low commercial and economical interest (Jacquemard et al., 1997; Corley & Law, 1997).

I.2.1.1.1. American oil palm *Elaeis oleifera* (HBK) Cortès

The American oil palm *E. oleifera* (HBK) Cortès, also known as *E. melanococca* (Hartley, 1988), has a distribution area going from Central America to the Amazon through Colombia and the Guyanas (Meunier & Boutin, 1975; Rajanaidu et al., 1986; Jacquemard et al., 1997; Corley & Tinker, 2016). *E. oleifera* palms are in general very small compared with their relative *E. guineensis*, with a procumbent stem although erected in some environment. *E. oleifera* as *E. guineensis*, is used domestically for the oil contained in its mesocarp and kernel. A significant proportion of its fruits develop in a parthenocarpic way, and the oil extracted from its pulp has a high content of unsaturated fatty acids, which gives it a fluidity comparable to that of olive oil (Meunier, 1969). Hybridization of *E. oleifera* with *E. guineensis* has been carried out and resulted to individuals with intermediate characteristics to their parents. However, this hybrid is economically of low interest given its partial sterility. Thereafter, oil palm will only refer to *E. guineensis*.

I.2.1.1.2. African oil palm, *Elaeis guineensis* Jacq.

Etymologically called olive tree of Guinea, oil palm (Fig. 1) originated from the Gulf of Guinea where its name comes from.



Fig. 1. Oil palm tree (Anonymous, 2020a).

It is a tree-like diploid with $2n = 2x = 32$ chromosomes, monocotyledon from the Arecaceae family (formerly called Palmae) (Jacquemard *et al.*, 1997).

The African origin of oil palm has long been controversial by the international scientific community until (Zeven, 1964) provides evidence showing an African origin. His work is based on the research of the first Botanists and the fossil pollen found in the soils of the Miocene in the Niger Delta (Fig. 2).

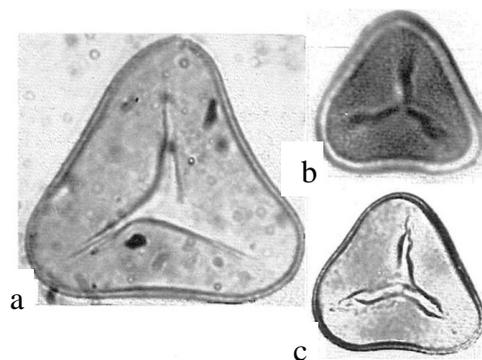


Fig. 2. Comparison of oil palm pollen (Nigeria). a: fossil pollen apparently similar to a fern spore (size $\times 1750$); b, c: fresh pollen of oil palm *pisifera* (size $\times 1300$) (Zeven, 1964).

I.2.1.2. Taxonomy and botanical description of *E. guineensis*

I.2.1.2.1. Taxonomy

The genus *Elaeis* belongs to the Arecaceae family, one of the oldest flowering plant families that exists, with fossils from the Cretaceous (Purseglove, 1976). *E. guineensis* belongs to the subfamily of Arecoideae containing approximately 60% of the genera of that family, therefore, 107 out of 183 and more than 50% of the species, i.e. approximately 1,300 out of 2,400 (Baker et al., 2011) making this subfamily the largest and most diverse of the five subfamilies of Arecaceae. Classification of *Elaeis* is made using the taxa below (Cronquist & Takhtadzhian, 1981; Dransfield et al., 2005; Corley & Tinker, 2016):

Domain: Eukaryota

Kingdom: Plantae

Subkingdom: Viridaeplantea

Phylum: Spermatophyta

Subphylum: Angiospermae

Class: Liliopsida/Monocotyledons

Order: Arecales

Family: Arecaceae/Palmae

Subfamily: Cocosideae/Arecoideae

Tribe: Cocoseae

Genus: *Elaeis*

Species: *Elaeis guineensis* Jacq. and *Elaeis oleifera* (HBK).

I.2.1.2.2. Botanical description

E. guineensis is a perennial tree plant with indefinite growth, presenting a crown extended from 30 to 45 green palms from 5 to 9 m long and a single cylindrical stipe (Rafflegeau, 2008). From 5 to 8 years old, its pinnate compound leaves bear 100 to 120 leaflets, while those at the base are transformed into thorns. The leaflets are quite short and about 7-10 cm wide. From 20 to 40 years old, a healthy tree has leaves that carry 190 to 200 leaflets from 70 to 90 cm long by 4 cm (sometimes 6 cm) wide and the petiole measures 70 cm to 1.10 m long and 25 cm wide (Chevalier, 1943).

The stipe or pseudo trunk of *E. guineensis*, has from 3 to 6 years a growth in length which goes from 30 to 75 cm per year. Its size can reach 25 to 30 cm long but its commercial exploitation stops when the tree exceeds 12 m. The diameter at the base is 80 to 110 cm, then

40 to 50 cm on the cylindrical area (Chevalier, 1943; Jacquemard, 2012). A lignified star-shaped cavity is present at the base of the bulb at the interface with the root system.

The oil palm's root system is made up of fasciculate adventitious roots originating on the root plate, which can reach 15 to 20 m in length and penetrate to around 6 m in depth. The voluminous root plateau of about 80 cm in diameter penetrates the soil to a depth of about 40 to 50 cm (Jourdan & Rey, 1997; Jacquemard, 2012).

I.2.1.3. Oil palm ecology

Oil palm is a plant that supports a very wide range of climatic factors. It is a plant at the edge of the forest and a gallery forest or shore (riversides). At the juvenile stage, it is sciaphile, therefore young plants usually need shade to resist drought in the savannah. As it develops, the need for light gradually increases, thus, becoming heliophile (Chevalier, 1943). Its cultivation is carried out in an interval of the humid tropical zone limited to 15° latitude on both sides of the equator (Henry, 1958; Jacquemard *et al.*, 1997). Maximum growth and production are obtained when the various climatic factors are at their optimum. Indeed, a minimum of 2,000 mm of precipitation well distributed i.e. without a pronounced dry season and ideally 100 mm at least each month is necessary throughout the year, the optimum insolation is beyond 1,800 hours (heliometers) and solar radiation above 12-15 MJ/m²/day and sunshine of 5-7 h/day (Hartley, 1988; Jacquemard, 1995, 2012; Goh, 2000). Maximum production is obtained for monthly average temperatures between 22 and 24°C. However, the monthly minima must be above 18°C and the maxima between 28 and 33°C because a blockage of bunches ripening and lethal effects occur if temperatures regularly drop below 18°C. Oil palm is not very demanding on its soil fertility and can therefore be cultivated on most tropical soils provided that they are deep, loose, not very grainy and well-drained (Hartley, 1988; Jacquemard *et al.*, 1997; Goh, 2000; Jacquemard, 2012).

I.2.1.5. Oil palm and environment

With an expected world population of over 9 billion by 2050, around 240 million tons of vegetable oil will be needed to supply the world demand, i.e., 120 to 156 million tons for palm oil (Corley, 2009; Rival & Levang, 2014). To supply palm oil, 12 to 28 million hectares of planted oil palm will be necessary depending on the performance of the planting material (Corley, 2009). It will therefore be necessary to increase the planted area and/ or the productivity of the already existing planted area. Oil palm is usually considered as a driver of deforestation (Butler *et al.*, 2009) and significant loss of animal biodiversity when a forest is

replaced with an oil palm plantation (Fitzherbert *et al.*, 2008). However, this belief can be misleading given the important part of destroyed forests not used for oil palm culture (Corley & Tinker, 2016). Indeed, from 1990 to 2000, around 78 million hectares of rainforest have been destroyed in the main 29 oil palm producers but the planted area of oil palm at the same period increased by only 3.9 million hectares (Anonymous, 2010), i.e., 5%. In consequence, deforestation due to oil palm culture accounted only for 5% of the total forest destroyed (Corley & Tinker, 2016). Moreover, from 2000 to 2010, 58 million hectares of forest were destroyed, while oil palm plantations expanded from 6 million hectares in the same 29 countries; corresponding to only 10% of deforestation (Anonymous, 2010).

I.2.1.4. Production and economic importance

Palm oil world production is distributed among many countries (Fig. 3). This production has increased steadily for the last 60 years. Starting with a production of around 1.5 Mt in the 1960s to over 75 Mt in 2020 (Anonymous, 2020c). This production is largely provided by two countries, Indonesia with 43.5 Mt and Malaysia with 19.3 Mt, i.e., 85% of the world production in 2020 for both. Cameroon, with a production estimated at 269,000 tons in 2018, although an increase compared to 2017, is still only the 13th world producer and fourth in Africa, behind Nigeria, Ivory Coast and Ghana. An important regression of Cameroon production compared to 2011 (354,000 tons) is acknowledged (Anonymous, 2020b). Although Indonesia and Malaysia are by far the largest producers of palm oil, India is the highest importer with 9.2 Mt, and Indonesia is the top domestic consumer with 14.875 Mt (Anonymous, 2020c).

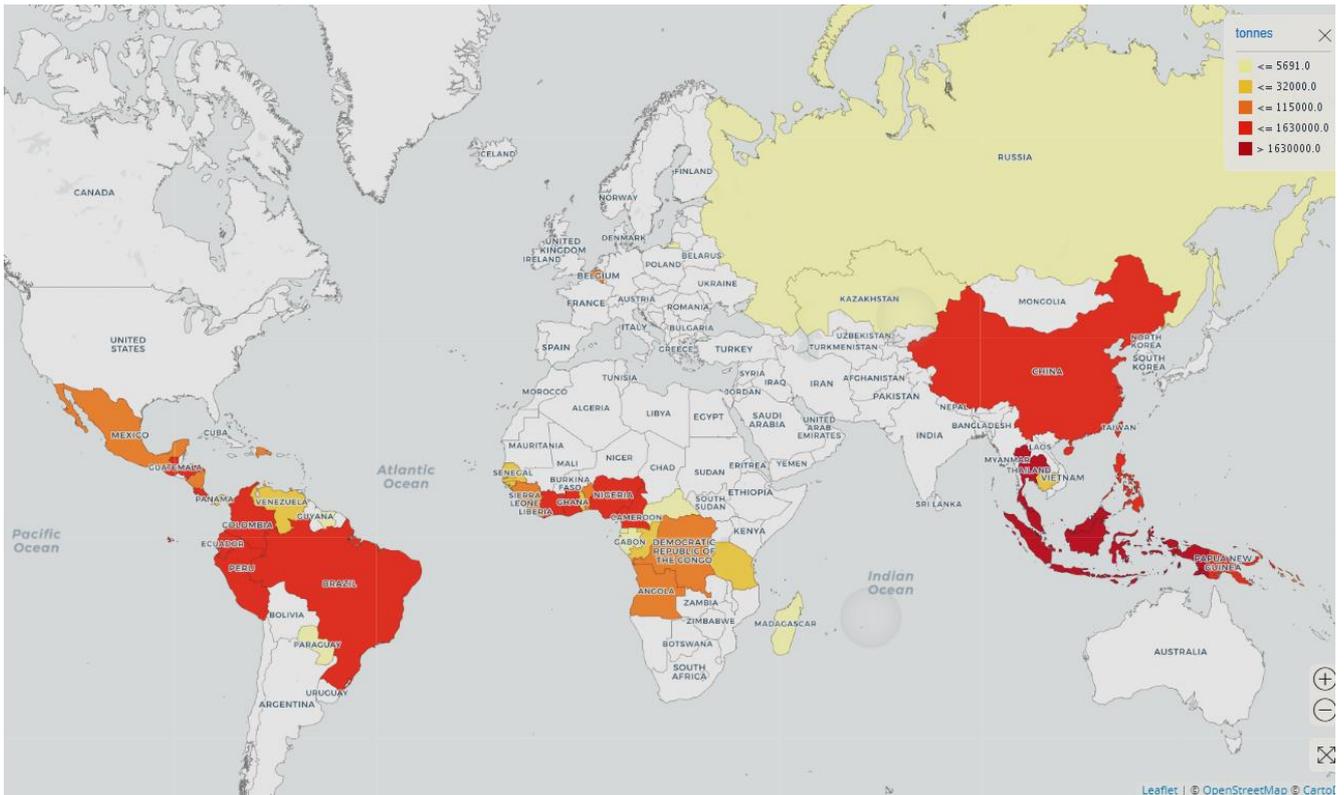


Fig. 3. Distribution of oil palm production worldwide (Anonymous, 2020b).

I.2.2. Concepts of quantitative genetics

I.2.2.1. Quantitative traits

Quantitative genetics is a special branch of genetics interested in the inheritance of quantitative traits i.e., traits jointly controlled by multiple genes of small effects and the environment. Phenotypic values of quantitative traits can vary in a range among individuals, thus giving a continuous distribution (Falconer & Mackay, 1996; Lynch & Walsh, 1998). Quantitative traits are therefore contrasted with Mendelian traits (also known as qualitative traits) whose phenotype is controlled by one or very few genes, with as consequence, a discrete distribution over individuals (Stearns, 1992). Genes responsible for quantitative traits are called quantitative traits loci (QTL) and usually, the segregation of these genes individually, expresses a small quantity of the genetic variance but collectively, a significant amount of the total genetic variance (Hayes & Goddard, 2001). In addition to the genetic factors, the phenotype over individuals can be explained by environmental factors and/ or their interaction with genetic factors, although the latter can be of less importance overall (Xu, 2013). Thanks to the progress of molecular biology, it becomes possible to link molecular markers to gene alleles, therefore, making a study of marker segregations possible whatever the gene's effect on the phenotype (Gallais, 2011). The variation of the phenotype due to individual QTL segregation effects are

usually difficult to observe, hence the necessity of appropriate statistical methods and mathematical models to value their effects, since most yield traits in cultivated plants are quantitative such as oil palm whose yield components are quantitative traits (Nyouma *et al.*, 2019).

I.2.2.2. Main properties of quantitative traits

Two main properties of quantitative traits are at the basis of breeding methods. First and foremost, the resemblance between relatives (explained below). The degree of resemblance between relatives varies across traits. Breeding strategies rely on the resemblance between parents and offspring, therefore, mating high-yielding parents will bring an improvement to the yield components of the next generation, depending on the degree of resemblance and their responsiveness to selection. The degree of resemblance between different relatives is used in breeding programs to predict the outcome of the breeding strategies in order to determine the best to be used (Falconer & Mackay, 1996).

The second property is the inbreeding depression. Indeed, this latter appears to diminish the mean of traits linked to fitness in animals and naturally outbreeding plants, thus leading to vigour and fertility losses. That loss is detrimental given that the majority of traits with high economic value for animals and plants are a feature of vigour or fertility. There are several techniques of inbreeding management mostly consisting of crossing in inbred lines (Falconer & Mackay, 1996).

I.2.2.3. Phenotypic value

Quantitative genetics focuses on genes involved in the expression of quantitative traits. In order to determine the link between the properties of a population as aforementioned and quantitative traits, the concept of phenotypic value should be introduced. This latter represents the first value obtained from the measure of quantitative traits. All the genetic parameters: population means, variance, covariance and heritability are derived from that value. The phenotypic value can be divided into components attributable to genetic and environmental (non-genetic) factors, and their interaction (Doolittle, 1987; Falconer & Mackay, 1996; Lynch & Walsh, 1998). The genetic components include the set of genes of an individual having an influence on the phenotype, while environmental components are non-genetic causes modifying the phenotype and their interaction. Hence, the phenotype results from genes' actions subsequently modified by environmental factors. Therefore, the basic model in quantitative genetics can be symbolically written as follow (Falconer & Mackay, 1996; Lynch & Walsh, 1998; Verrier *et al.*, 2001):

$$P = G + E \quad [1.1]$$

with P the phenotypic value of the population, G the genetic (genotypic) value and E the environmental deviation or environmental value.

It is convenient for an individual or a group of individuals to express the phenotypic value in terms of deviation from the population mean. Hence, a more useful form for expressing the phenotypic value is (Doolittle, 1987; Verrier *et al.*, 2001):

$$P_i = \mu + G_i + E_i \quad [1.2]$$

with μ the phenotypic population mean, G_i the genetic value of the individual i and E_i the environmental effects on the individual i .

From equation [1.2], it is possible to obtain:

$P = \mu + A_i + D_i + I_i + E_i$, with A_i the additive genetic value of the individual i and D_i the dominance genetic value or dominance deviation of the individual i and I_i the epistatic genetic value of the individual i , $G_i = A_i + D_i + I_i$. When D_i and I_i are negligible only the additive effects remain.

1.2.2.3.1. Genotypic values at one diallelic locus

Since the deviation due to environmental factors is not negligible, measuring genotypic value is feasible only theoretically but impossible in practice, unless if one locus only is involved with genotypes resulting to distinct phenotypes or in genotypes of high inbred lines (Falconer & Mackay, 1996).

Consider a locus A with two alleles, A_1 the allele that increases the genotypic value and A_2 the allele that reduces it. Let $+a$ be the genotypic value of the homozygote A_1A_1 , $-a$ the genotypic value of the second homozygote A_2A_2 and d the genotypic value of the heterozygote A_1A_2 (Fig. 4).

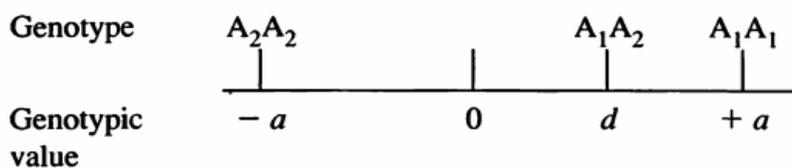


Fig. 4. Genotypic value based on one locus-genotype, with randomly assigned alleles (Falconer & Mackay, 1996; Lynch & Walsh, 1998; Conner & Hartl, 2004).

The midpoint between the two homozygotes is 0. Heterosis or hybrid vigor of the heterozygote depends on the values taken by d i.e., the dominance degree. For $d = 0$, there is

no dominance; for $d > 0$, A_1 is dominant over A_2 , for $d < 0$ A_1 is dominant over A_2 . In case of overdominance ($A_1A_2 > A_1A_1$) $d > +a$, and complete dominance if $d = +a$ ($A_1A_2 = A_1A_1$, with A_1 being dominant over A_2) or $d = -a$ ($A_1A_2 = A_2A_2$, with A_2 being dominant over A_1). The ratio d/a allows expressing the degree of dominance (Falconer & Mackay, 1996; Lynch & Walsh, 1998; Conner & Hartl, 2004; Gallais, 2011).

I.2.2.3.2. Genotypic mean of a population

The genotypic mean of the population when the population allele (or gene) frequencies of A_1 and A_2 are p and q respectively, can be computed as follow (Falconer & Mackay, 1996; Conner & Hartl, 2004):

$M = ap^2 + 2dpq + (-a) \times q^2$, with p^2a the mean of A_1A_1 , $2pqd$ the mean of A_1A_2 and $-q^2a$ the mean of A_2A_2 (Table I).

$$M = a(p - q) + 2dpq \quad [2]$$

In case there are many loci involved as in quantitative traits, and it is assumed that genes additionally combine i.e., the genotypic value is the sum of values of each locus taken independently. The equation is expressed as follow:

$$M = \sum[a(p - q) + 2dpq] \quad [3]$$

Table I. Deduction of the population genotypic mean from the relative allele frequencies and genotypic value (Falconer & Mackay, 1996; Conner & Hartl, 2004).

Genotype	Frequency	Genotypic value	Frequency \times Genotypic value
A_1A_1	p^2	$+a$	p^2a
A_1A_2	$2pq$	d	$2pqd$
A_2A_2	q^2	$-a$	$-q^2a$
	Sum		$a(p - q) + 2pqd$

I.2.2.3.3. Average effect of allele substitution

Once seen how the genotypic mean can be calculated in a population, the following step is the understanding of the transmission of genes from parents to their progenies. The knowledge of the genotypic mean does not provide such information given that genotypes are made up in each generation. The average allele (gene) effect can be defined as the average deviation from the population mean of individuals that received a given allele from one parent, with the allele of the other parent assumed to come randomly from the population. In other

words, if 10 individuals carrying the allele are combined with random alleles in that population, then the deviation of the mean genotype obtained from the population mean is the average effect (Falconer & Mackay, 1996; Gallais, 2011).

In order to link the genotypic mean and the average effect, two alleles A_1 with a frequency p and A_2 with a frequency q are considered. Let α_1 the average effect of A_1 . If the gametes carrying A_1 randomly unite with gametes from the population, the frequencies of the genotypes involving A_1 will be p of A_1A_1 , q of A_1A_2 , with a mean of $pa + qd$ (Fig. 5). The average effect α_1 of A_1 is the difference between this mean and the population mean calculated above. Simplification enables to obtain (Fisher, 1918; Falconer & Mackay, 1996):

$$\alpha_1 = pa + qd - [a(p - q) + 2dpq]$$

$$\alpha_1 = q[a + d(q - p)] \quad [4]$$

The average effect of A_2 is computed similarly as:

$$\alpha_2 = -p[a + d(q - p)] \quad [5]$$

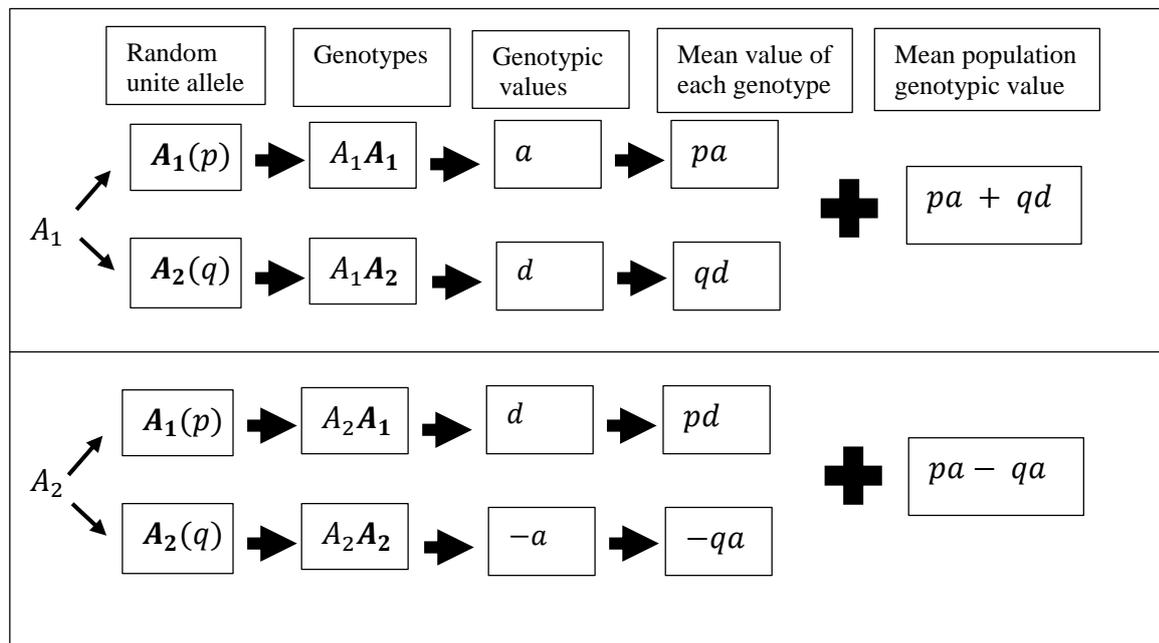


Fig. 5. Allele substitution and genotypic values of the resulting genotypes.

It is recommended to express the average effect in terms of the average effect of allele substitution. That corresponds when loci involve only two alleles, to the difference between the average effect of the two alleles A_1 and A_2 :

$$\alpha = \alpha_1 - \alpha_2$$

$$\alpha = a + d(p - q) \quad [6]$$

The average effects of A_1 and A_2 in terms of the average effect of the allele substitution, are:

$$\left. \begin{aligned} \alpha_1 &= q\alpha \\ \alpha_2 &= -p\alpha \end{aligned} \right\} [7]$$

I.2.2.3.4. Breeding value or additive genetic value

The breeding value of an individual is the value passed on average to its progeny. Given that parents transmit their alleles and not their genotypes to their progeny, the breeding value can therefore be computed as the sum of allele average effects for all the loci. While average effects cannot be measured, the breeding values can, through its progeny. Indeed, if an individual is randomly mated with several other random individuals, its breeding value is twice the mean deviation of its offspring from the population mean. One parent passes on only half of its genes, hence the deviation is doubled (Falconer & Mackay, 1996; Conner & Hartl, 2004; Gallais, 2011). The breeding value is a function of the individual and the population in which its mates are randomly drawn. The breeding value can be measured for traits that one parent does not possess. An example to illustrate that is the breeding value of a bull for milk production, although does not produce milk strictly speaking. This can be done on its offspring in which the measures are done (Conner & Hartl, 2004). Considering [7], the different breeding values will be (see Doolittle (1987); Falconer & Mackay (1996)):

$$\alpha_1 = 2q\alpha \text{ for } A_1A_1 \quad [8.1]$$

$$\alpha_2 = -2p\alpha \text{ for } A_2A_2 \quad [8.2]$$

$$\alpha_1 + \alpha_2 = (q - p)\alpha \text{ for } A_1A_2 \quad [8.3]$$

The population means i.e., the means including all the three genotypes is obtained by summing their respective breeding values. Following that reasoning, the mean population (M) is:

$M = 2q\alpha - 2p\alpha + (q - p)\alpha$, by substituting α by its value $a + d(p - q)$, we find:

$M = a(p - q) + 2dpq$ which is the expression of breeding value without average effects.

The expected breeding value of a given individual is the average breeding value of its two parents. As a consequence, different descendants of the same parents can have different breeding values depending on the received alleles from their parents. The solution to deal with that is to calculate the expected breeding value in a large number of descendants of the same parent as (Falconer & Mackay, 1996):

$$A_i = \frac{1}{2}(A_{p_m} + A_{p_f}) \quad [9]$$

with A_i the expected breeding value of an individual i , A_{p_m} and A_{p_f} the breeding values of the male (p_m) and female (p_m) parents of i .

Similarly, the dominance deviation or dominance value can be expressed in terms of assigned genotypic values a and d . Therefore, the genotypic values should be converted into deviation to the population given that breeding values have already been expressed that way. To illustrate that, consider A_1A_1 with its assigned genotypic value $+a$. The genotypic value $+a$ in the form of deviation to the population can be obtained as the difference between the genotypic value, $+a$ and the mean population genotypic value (M) as follow (Falconer & Mackay, 1996):

$$\begin{aligned} a - M &= a - [a(p - q)2dpq] \\ &= a(1 - p + q) - 2dpq \\ &= 2q(a - dp) \quad [10a] \end{aligned}$$

That equation can be expressed with average effects by replacing a by its value $\alpha - d(q - p)$; thus becoming:

$$2q(\alpha - qd) \quad [10b]$$

The dominance deviation is finally obtained by subtracting the genotypic value of A_1A_1 in [10b] by its breeding value, $2q\alpha$ in [8.1].

$$2q(\alpha - qd) - 2q\alpha = -2q^2d \quad [11].$$

Similarly, the dominance deviation of A_1A_2 and A_2A_2 can be obtained as: $2pqd$ and $-2p^2d$, respectively.

I.2.2.3.5. Genetic value

The difference between breeding value and genetic or genotypic value is dominance deviations indicated by vertical dotted lines (Fig. 6) and epistasis deviations. The genetic value can, therefore, be divided into two parts: additive genetic value (breeding value) and non-additive genetic value (dominance and epistasis) observable on the individual itself. The dominance genetic value results from an interaction of alleles within a locus while epistatic value results from an interaction of alleles from different loci. Dominance genetic effect is the most important non-additive genetic effect (Falconer & Mackay, 1996; Gengler *et al.*, 1998). The link between genotypic values, breeding values and dominance deviation is illustrated in

Fig. 6. The genotypic values are represented in relation to the number of A_1 alleles present in the genotype. A regression line is fitted by points and each point is weighted by the frequency of its genotype. The line provides the breeding values of each genotype and the upper cross mark on it is the population mean. The average allele effect is the allele substitution α , the difference between A_2A_2 and A_1A_2 or A_1A_2 and A_1A_1 (Falconer & Mackay, 1996).

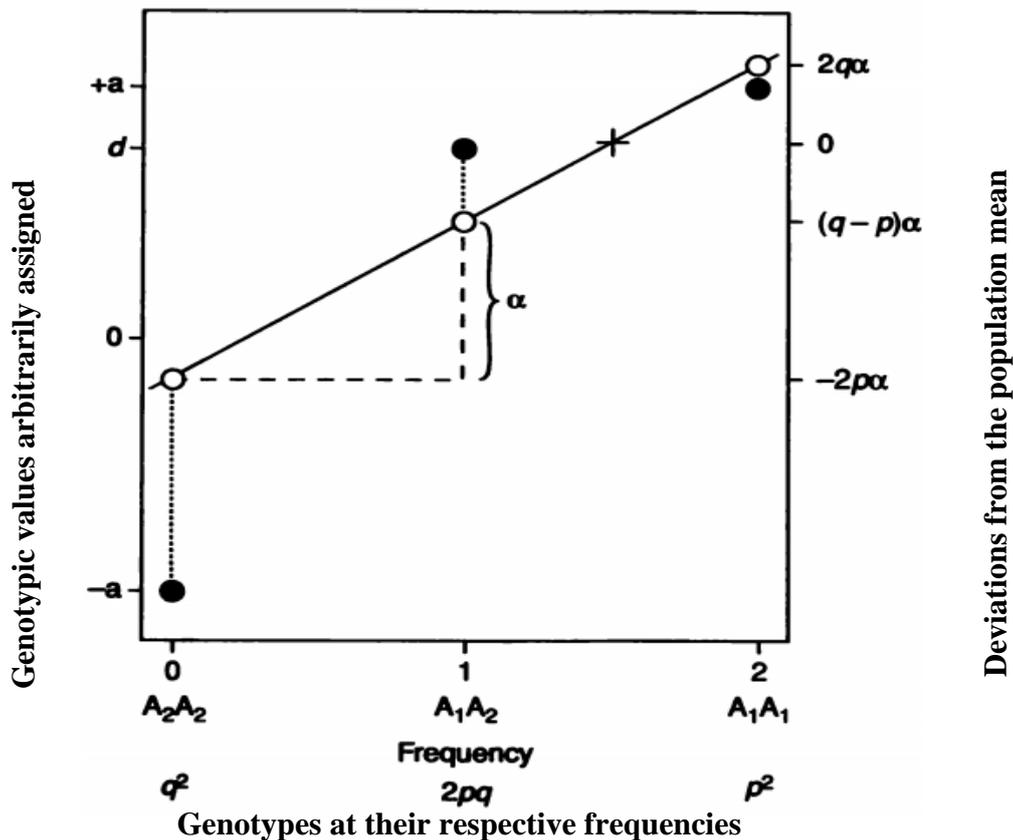


Fig. 6. Decomposition of phenotypes into genotypic values (closed circles), breeding values (open circles), for a locus with two alleles, A_1 and A_2 at frequencies p and q . $d = \frac{3}{4}a$ and $q = \frac{1}{4}$ and α is the average effect of allele substitution (Falconer & Mackay, 1996).

I.2.2.4. Phenotypic variance

I.2.2.4.1. Definition of the components of phenotypic variance

As the population mean aforementioned, variance is an important parameter for the characterization of quantitative traits in a population. Variance is a fundamental statistical measure of the amount of variation from which other parameters and tests are based (Conner & Hartl, 2004). Variance is also the mean of the square deviation of a random variable from its mean or population mean. In plant breeding, the total variance corresponds to the phenotypic variance, also known as the variance of phenotypic values, and can be computed by summing separately all its components. Indeed, the phenotypic variance can be partitioned into variances

of the phenotypic components, total genotypic variance (V_G), environmental variance (V_E) and their interactions, assuming there is no interaction or correlation between genetic and environmental factors (Gallais, 2011):

$$V_P = V_G + V_E + V_{G \times E} \quad [12]$$

For a better understanding of the phenotypic variance, some basic concepts mentioned in previous paragraphs should be known, among which the population mean, the average (genetic) effect and the breeding value.

Genetic variance often termed genotypic variance can be fragmented into additive and non-additive genetic variances. In a given population, the additive genetic variance expresses the variance of additive effects of genes, i.e., the sum of additive effects at each locus. Dominance genetic variance is the sum of statistical dominance variance at each locus. In the absence of epistatic effect, total genetic variance is the sum of dominance and additive genetic variances when the population is in Hardy-Weinberg equilibrium (Falconer & Mackay, 1996; Gallais, 2011): $V_G = V_A + V_D + V_I$

Hence, [12] becomes: $V_P = V_A + V_D + V_I + V_E + V_{G \times E}$.

Often, $V_{G \times E}$ can be neglected without significantly affecting the phenotypic variance (Falconer & Mackay, 1996).

I.2.2.4.2. Additive genetic variance

The additive genetic variance is the only genotypic variances that can be estimated from field observations of the population. Response to selection of the population is usually proportional to the genetic (additive) variance (Toro *et al.*, 2011), hence its importance in plant breeding. To determine the additive variance in practice, the total variance is partitioned in additive variance against all the other forms of variances. Additive variance does not mean alleles or genes act additively with non-additive actions (dominance and epistasis). Additive variance in the scale of locus is the average effect of its different alleles (Kempthorne, 1955) and no assumption should be made on gene action modes.

Consider a single locus with two alleles (excluding within-loci interactions), and let express the genetic variance in the form of gene frequencies (p and q) and genotypic value (a and d). The additive genetic variance corresponding to the variance of breeding value can be calculated by multiplying the squared breeding value ([8.1], [8.2], [8.3]) of each genotype by its frequency and summing as follow (Falconer & Mackay, 1996; Conner & Hartl, 2004):

$$V_A = p^2(2q\alpha)^2 + 2pq(q-p)^2\alpha^2 + (-2p\alpha)q^2$$

$$V_A = 4p^2q^2\alpha^2 + 2pq(q-p)^2\alpha^2 + 4p^2q^2\alpha^2 \quad [13.1]$$

$$V_A = 2pq\alpha^2 \quad [13.2] \text{ (expressed in terms of average effect)}$$

$$V_A = 2pq[a + d(q-p)]^2 \quad [13.3] \text{ (expressed in terms of assigned genotypic values } a \text{ and } d).$$

I.2.2.4.3. Dominance genetic variance

The dominance variance can be expressed similarly to the genetic additive variance as follow (Falconer & Mackay, 1996; Conner & Hartl, 2004):

$$V_D = p^2(-2q^2d)^2 + q^2(-2p^2d)^2 + 2pq(2pqd)^2$$

$$V_D = (2pqd)^2 \quad [14]$$

Overall, all the variance components have a squared term because variance has previously been defined as a square deviation from the population mean. That term prevents variance components from being negative because negative genetic variability is meaningless except in practice where it often occurs due to random error (Conner & Hartl, 2004).

I.2.2.4.4. Total genetic variance without epistasis

Assuming epistatic effects are negligible, the total genetic variance can be expressed as:

$$V_G = V_A + V_D + 2Cov(A, D) \quad [15]$$

$Cov(A, D)$ being the covariance of breeding value and dominance deviation. $Cov(A, D)$ can be calculated as the sum of the product of breeding value by dominance deviation and the frequency, of each genotype. Thus, it can easily be demonstrated that $Cov(A, D)=0$, hence

$$V_G = V_A + V_D$$

$$V_G = (2pqd)^2 + 2pq[a + d(q-p)]^2 \quad [16]$$

In case of absence of dominance ($d = 0$),

$$V_G = V_A = 2pq\alpha^2 \quad [17]$$

I.2.2.4.5. Total genetic variance with epistasis

Two considerations of epistasis phenomenon as any type of genetic effects are possible: the physiological or biological epistasis and the statistical epistasis. Epistasis can biologically be defined as a phenomenon in which the phenotype of an individual with several genotypes at one locus depends on the genotypes at the other loci (Cheverud & Routman, 1995). Statistically,

epistasis will be defined as already mentioned i.e. the deviation of the genotypic values of many loci from the expected value based on the sum of the value of each locus (Falconer & Mackay, 1996). Just like all the statistical parameters in quantitative genetics, statistical epistasis is function to the population and the allele frequency, while biological epistasis depends on the individual genotype, and is independent of the population and remains constant, even if the allele frequency changes (Cheverud & Routman, 1995; Goodnight, 2016). Hereafter, epistasis will refer to statistical epistasis.

When at least two loci are involved, from their interactions arises epistatic variance, called by some authors, variance interaction deviations (V_I) (Falconer & Mackay, 1996). Epistatic variance can be theoretically explained depending on the number of loci involved on one hand, and the type of genetic effects (breeding or dominance genetic value) on the other hand.

Firstly, the number of loci involved is proportional to the number of factors involved in the interaction; for instance, between two loci, two interaction factors will be involved and so on. Moreover, when a large number of loci are implicated, there is also a large number of interaction factors so that, the epistatic variance is minimized and negligible.

Secondly, when considering breeding and dominance values, three forms of epistasis are possible: additive interaction at both loci, additive interaction at one locus and dominance at the other and dominance at both loci leading respectively to additive \times additive variance ($V_{I_{A \times A}}$), additive \times dominance variance ($V_{I_{A \times D}}$), dominance \times dominance variance ($V_{I_{D \times D}}$) (Falconer & Mackay, 1996). As a result, epistasis variance for two interaction factors is expressed as follow: $V_I = V_{I_{A \times A}} + V_{I_{A \times D}} + V_{I_{D \times D}}$. In many studies, epistasis variance shown to be non-significantly different from zero i.e., negligible over all the phenotypic variance (Su *et al.*, 2012) and therefore, will be ignored here.

Estimation of the genetic variance is just theoretical, to the extent that in practice, gene frequencies and gene effects are unknown, unless if a special population is made up accordingly (Falconer & Mackay, 1996). In practice, genetic variance is estimated through its components, and easily when data about relative resemblance are available.

1.2.2.5. Resemblance between relatives

The theoretical resemblance between relatives caused by genetic factors was first ascertained by Fisher (1918). Based on his theory, quantitative genetic factors are estimated thanks to the resemblance between different types of relatives by linking phenotypic covariance to the degree of genetic relationship, usually expressed as the kinship coefficient (or coancestry

coefficient) (Lynch & Walsh, 1998) or the additive coefficient of relationship (Falconer & Mackay, 1996; Lynch & Walsh, 1998).

I.2.2.5.1. Genetic relationships

Kinship coefficient or coancestry coefficient (Wright, 1922; Malécot, 1948) can be computed using the pedigree—genetic genealogical relationship or using molecular markers—genetic realized or molecular relationship, which can be either additive or nonadditive (dominance or epistatic genetic relationship) (Visscher *et al.*, 2006). Indeed, the coefficient of kinship (f_{ij}) is a probabilistic measure of relatedness or relationship between two individuals i and j , defined as the probability that a pair of homologous alleles randomly sampled at a given locus are identical by descent (IBD). In other words, it is the probability to have for the same locus the allele of an individual i identical to the allele of an individual j , and coming from a common recent ancestor. It ranges from 0 – 1, with 0 corresponding to unrelated individuals and 1 corresponding to pure lineage (Table II).

Two alleles are IBD (Fig. 7) if they are homologous alleles inherited from a common recent ancestor. Two alleles can be identical by state (IBS) i.e., alleles that are identical or similar regardless of whether they are inherited from a common ancestor. Therefore, IBD alleles or genes are IBS but not conversely (Lange, 2003; Powell *et al.*, 2010). The concept of IBD must be defined for a given base or reference population. In other words, the probability to have the same allele inherited from a common recent ancestor must be applicable only if the two individuals belong to the same studied reference population.

Table II. Kinship and fraternity coefficients according to their family relationship.

Individual relationships	Kinship coefficient (f_{ij})	Fraternity coefficient (ϕ)
unrelated individuals	0	0
pure lineage	1	1
individual - self	1/2	1
clones	1/2	1
fullsibs (with unrelated parents)	1/4	0.25
parents - offspring	1/4	0
grandparents-grandchild	1/8	0
half-sib	1/8	0
great grandparent - great grandchild	1/16	0

Classically, IBD is computed using a pedigree spanning many generations, in which the first individuals (at the top) are considered to be the founders (with no known parents) and assumed unrelated and noninbred. With the advent of molecular biology, it is now common to compute IBD using molecular markers (SSRs or SNPs) (Powell *et al.*, 2010).

An individual is inbred if both of its two parents are related. The consequence of inbreeding is the possibility for an individual to have received for a given locus two allele copies (IBD) of the same gene present in the common ancestor of its two parents (Fig. 8).

The coefficient of inbreeding of an individual x , F_x , is the probability to have two identical copies of the same allele (IBD) in a given locus. The two alleles of x being the result of a random draw of a gene among the two of its male parent and of a gene among the two of its female parent, the coefficient of inbreeding of an individual is equal to the coefficient of kinship between his two parents i.e. x with female parent i and male parent j ; we obtain (Wright, 1922): $F_x = f_{ij}$

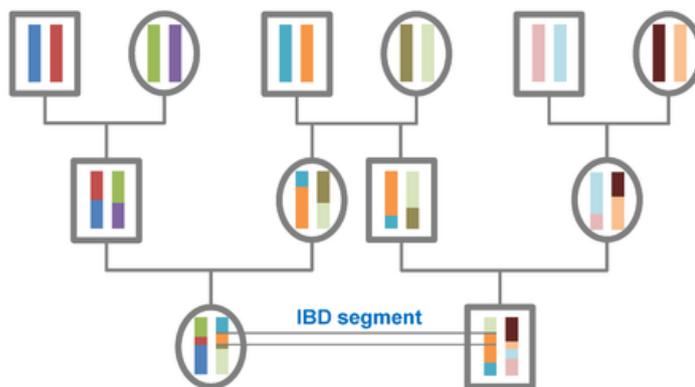


Fig. 7. Transmission of identical by descent segment of chromosome in two offspring (IBD) (Anonymous, 2013).

Kinship and inbreeding concepts are often confounded, whereas, although close, they are quite different. While kinship concerns pairs of individuals, inbreeding involves single individuals. Confusion commonly occurs because in common parlance, consanguineous refers to the fact of descending from the same strain. However, geneticists refer to marriage between relatives and reserve the term consanguineous for children born of such a marriage (Verrier et al., 2001).

To know how f_{ij} is computed, the estimation of the coefficient of coancestry of an individual with itself (self-coancestry), f_{xx} is necessary. Let us assume that an individual x carries in a given locus two alleles x_1 and x_2 . Now, consider that two alleles of x are randomly drawn in that locus. Given that f_{xx} is the probability two alleles are IBD, there are four possibilities, with a probability of $\frac{1}{4}$ each: x_1x_1 (IBD), x_2x_2 (IBD), x_1x_2 (non-IBD), x_2x_1 (non-IBD). Consequently, $f_{xx} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. However, individual x could be inbred; in that case F_x is the probability that x_1 and x_2 are IBD. In result, the coefficient of self-coancestry becomes

$f_{xx} = \frac{1}{4} + \frac{1}{4} + \frac{1}{4}F_x + \frac{1}{4}F_x$. By simplifying this equation, we obtain: $f_{xx} = \frac{1}{2}(1 + F_x)$ (Falconer & Mackay, 1996; Verrier *et al.*, 2001).

The coefficient of coancestry between a parent and its offspring can be obtained quite similarly, although slightly more complicated. Let us consider now two unrelated parents M and P , with a pair of alleles m_1 and m_2 , p_1 and p_2 , respectively. Four different types of descendants can be obtained: m_1p_1 , m_1p_2 , m_2p_1 , m_2p_2 . To make this example simple, let us compute the coancestry coefficient between the parent m_1m_2 and its offspring m_1p_1 . Here, between m_1m_2 and m_1p_1 , there is only one possibility (probability equal to $\frac{1}{4}$) to obtain IBD alleles (m_1m_1) among the four. This coancestry coefficient is exactly the same whichever offspring individual or parent taken. To summarize, the coancestry coefficient between a parent (unrelated to the second parent) and its offspring (non-inbred) is $\frac{1}{4}$. However, this conclusion assumes that the two parents are unrelated and the offspring is non-inbred. In case the two parents are related this coancestry coefficient will increase (Falconer & Mackay, 1996). To sum up, when literature says the coefficient of coancestry between a parent and its offspring is $\frac{1}{4}$, it implies that the parents are not related, and the offspring is not inbred as well.

Moreover, the coancestry coefficients between two individuals x and y (f_{xy}) can also be calculated between full sibs or in more complex relatedness schemes using the generalizing formula (Boucher, 1988; Lynch & Walsh, 1998):

$f_{xy} = \sum_i f_{ii} \left(\frac{1}{2}\right)^{n_i-1} + \sum_j \sum_{j \neq k} f_{jk} \left(\frac{1}{2}\right)^{n_{jk}-2}$, with n_i the number of individuals in the path (x and y included) leading to i the common ancestor, n_{jk} the number of individuals in the path conducting to j and k the two related but different ancestors (Lynch & Walsh, 1998).

Inbreeding coefficient is defined for a given neutral locus; therefore, their values depend on the length and the reliability of the pedigree. Inbreeding coefficients range also from 0 to 1.

The double of the kinship coefficient termed relationship or relatedness coefficient is used in practice to elaborate the additive relations relationship matrix, which is also referred to as the numerator relationship matrix (Lynch & Walsh, 1998). A relationship matrix is a square matrix with the same individuals in rows and columns, giving the self-relationship coefficients on the diagonal and relationship coefficients between distinct individuals off-diagonal. These relationships matrices are used for predictions purposes of the general combining ability (GCA) or additive genetic values.

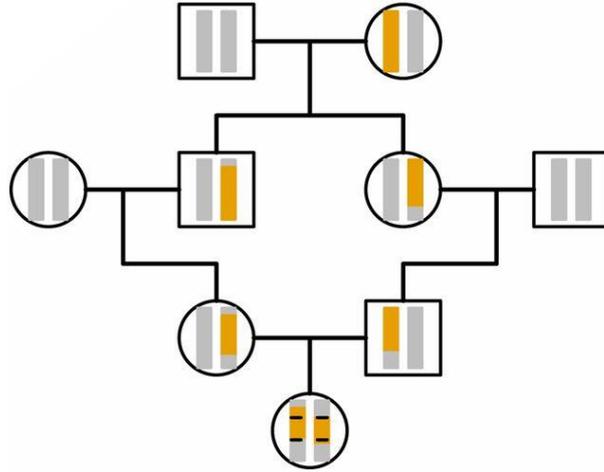


Fig. 8. Inheritance of two identical segments from a common ancestor in an inbred individual (Severson et al., 2019).

The coefficient of fraternity is another useful measure of the resemblance between relatives, defined as the probability for both alleles of a given locus in a pair of individuals to be IBD (Trustrum & Williamson, 1961; Lynch & Walsh, 1998). When considering two individuals x and y with their male parents p_x and p_y , and their female parents m_x and m_y , respectively, there are ways to draw a pair of IBD alleles between x and y . Firstly, the couple of alleles of p_x and m_x can be IBD and that of p_y and m_y maybe IBD. Secondly, the couple alleles from p_x maybe be IBD with that of m_y , and that from p_y can be IBD with that from m_x . Hence the following formula (Lynch & Walsh, 1998):

$$\varphi_{xy} = f_{p_x p_y} f_{m_x m_y} + f_{p_x m_y} f_{p_y m_x} \quad [18]$$

If x and y are full sibs [18] becomes:

$$\varphi_{xy} = f_{pp} f_{mm} + f_{pm} f_{pm} \quad [19]$$

If in addition p and m are unrelated, $f_{pp} = f_{mm} = \frac{1}{2}$ and $f_{pm} = 0$, with as result $\varphi_{xy} = 1/4$ (Table II). In case one of the two parents of [18] are not related, $\varphi_{xy} = 0$.

The fraternity coefficient is used to calculate the dominance relationship matrix used in the prediction of specific combining ability (dominance) or non-additive genetic value (here dominance genetic value).

I.2.2.5.2. Genetic covariances between relatives

The merit for clarifying the link between phenotypic resemblance and genetic variances in populations goes to Fisher (1918), Wright (1921), Cockerham (1954) and Kempthorne (1954). Statistical methods using the maximum likelihood and software programs have been

developed to compute genetic variance and covariance (Gilmour et al., 1995; Neale et al., 2003). These methods are able to estimate variance components using observed variations between and within families (Falconer & Mackay, 1996). Phenotypic variance, as aforementioned, can be partitioned into environmental and genetic components. Assuming that genetic and environmental components are not correlated when the genotypes are distributed in different environments, the covariance between two phenotypes of individuals i and j can be expressed as (Fisher, 1918):

$Cov(P_i, P_j) = Cov(G_i, G_j) + Cov(E_i, E_j)$. If i and j are drawn randomly and independently (not related), $Cov(G_i, G_j) = 0$. $Cov(G_i, G_j) \neq 0$ if i and j are related (not taken independently) i.e., have a common ancestor (have alleles IBD). If the individuals are in different environments, $Cov(E_i, E_j) = 0$. When individuals do not have a common environment, phenotypic covariance comes down to genotypic variance: $Cov(P_i, P_j) = Cov(G_i, G_j)$.

It is fundamental to identify cases where the common environmental factors are significant. That situation is observed when uncontrolled environmental factors, or whose effect cannot be corrected, affects several individuals (Falconer & Mackay, 1996).

Genetic covariance can be partitioned into additive and non-additive components. The latter is divided into dominance and epistasis components. Here, epistasis will be considered negligible. Assuming that variables A and D are not correlated i.e., independent, the phenotypic covariance between the individuals becomes:

$$Cov(P_i, P_j) = Cov(A_i, A_j) + Cov(D_i, D_j) \quad [20]$$

The additive covariance is non-null, if the two individuals have IBD alleles. For the dominance covariance to be non-null, the two individuals must have received, from their two respective parents, the same pair of genes; in other words, their coefficient of fraternity should be non-null. To summarize, genetic covariance is therefore non-null if the two individuals have received each at least one copy of the same gene present in a common ancestor. The calculation of the covariance between relatives involves the probabilities of identity of the genes (Fisher, 1918; Malécot, 1948). We can easily demonstrate that [20] becomes (Falconer & Mackay, 1996):

$$Cov(G_i, G_j) = 2f_{xy}V_A + (f_{p_x p_y} f_{m_x m_y} + f_{p_x m_y} f_{p_y m_x})V_D$$

$$Cov(G_i, G_j) = 2f_{ij}V_A + \varphi_{ij}V_D$$

I.2.2.5.3. Concept of heritability

Environment sometimes can influence and disturb the correspondence between the expected phenotypic value based on gene effects and the phenotypic value obtained (Wray & Visscher, 2008; Gallais, 2011). Heritability is the degree of expression of genetic factors on phenotypes; in other words, the amount of phenotypic variance due to genetic causes. It can be used also to designate resemblance between parents and their offspring (Wray & Visscher, 2008). Heritability is one of the most important properties of quantitative traits mostly due to its ability to show the evolution of phenotypes in response to selection (natural or artificial) (Conner & Hartl, 2004). The values of heritability can range from 0 (if the total variation is due to environmental causes) to 1 when the total variation is due to genetic causes (Corley & Tinker, 2016).

I.2.2.5.2.1. Broad sense heritability

Broad sense heritability (H^2) is the proportion of phenotypic variance that is of genetic origin. In other words, it is the ratio between the genetic variance and the phenotypic variance (Verrier et al., 2001; Conner & Hartl, 2004; Gallais, 2011). Broad sense heritability includes variance due to dominance and epistasis factors, therefore, is more useful in clonal selection and in the selection of highly self-fertilizing species whose genotypes are almost intactly passed on from parents to offspring (Conner & Hartl, 2004). The broad-sense heritability is expressed as follow (Verrier et al., 2001; Conner & Hartl, 2004; Gallais, 2011):

$$H^2 = \frac{V_G}{V_P} = \frac{V_A + V_D + V_I}{V_A + V_D + V_I + V_E} \quad [21]$$

I.2.2.5.2.2. Narrow-sense heritability

Narrow-sense heritability or *sensu stricto* heritability (h^2) is the ratio between the genetic additive variance and the phenotypic variance. Narrow-sense heritability is mostly useful in the selection of outbreeding species (Conner & Hartl, 2004) such as oil palm. Narrow-sense heritability is expressed as (Conner & Hartl, 2004; Gallais, 2011):

$$h^2 = \frac{V_A}{V_P} = \frac{V_A}{V_A + V_D + V_I + V_E} \quad [22]$$

I.2.3. Overview of oil palm genetics and breeding strategies

I.2.3.1. Oil palm breeding goals and objectives

A breeding goal is a direction to follow in the improvement of traits of interest including the emphasis of each trait of a crop population. A breeding goal usually focused on economical

profit although the quality of the product is also taken into consideration. Afterwards, breeding objectives are defined based on traits on which breeding should be oriented in order to make the culture economically profitable.

The breeding goal in oil palm is to increase the yield and make an economically profitable oil palm culture. To achieve that, many objectives are set by research programs, among which priority is respectively given to agronomic traits: oil yield increase, disease resistance (among which Ganoderma basal stem rot and Fusarium wilt, Crown disease) and high bunch index or yield. As suggested by Corley (2009, 2006), palm oil yield per hectare could reach up to 18 tons of oil per year if growth and yield components are at their optimum. The other traits in oil palm breeding programs are to simplify the harvesting process: slow height increase, long bunch stalk, oil composition (low lipase, high oleic acid, carotene content), stress tolerance (drought tolerance, low-temperature tolerance, etc.) (Jacquemard et al., 1997; Corley & Tinker, 2016; Soh et al., 2017).

I.2.3.2. Genetic determinism and fruit forms

The understanding of the genetic determinism of the fruit form was acquired in the 1930s (Beirnaert & Vanderweyen, 1941). Fruit form is genetically controlled by a gene, now named *SHELL* (*Sh*), with two codominant alleles Sh^- and Sh^+ at the origin of three fruit form in oil palm (Fig. 9.). *pisifera* $Sh^-//Sh^-$ and *dura* $Sh^+//Sh^+$ are thus homozygotes and *tenera* $Sh^+//Sh^-$ heterozygote. *pisifera* is a shell-less natural mutant usually female sterile with lignified fiber pulp, naturally present in nature at less than 0.5%. *dura*, has a thick-shell greater than 2 mm and therefore a small pulp (or mesocarp) quantity, and a mesocarp ranging from 2 to 6 mm size and taking 35–65% of fruit quantity. *dura* is the most abundant in spontaneous and sub-spontaneous palm groves i.e., around 97% of the total palms. *tenera* is the hybrid of the cross between *dura* and *pisifera* with a thin shell lesser than 2 mm and a ring of lignified fibers in the pulp around the kernel (Cochard et al., 2001; Demol, 2002; Corley & Tinker, 2016). The form cultivated in commercial plantations since the 1950s is *tenera*, as it combines a high percentage of pulp per fruit (PF) with female fertility, and is obtained by the cross *dura* × *pisifera*. Its use instead of the traditional *dura* increased oil palm yield by 30% (Corley & Tinker, 2016).

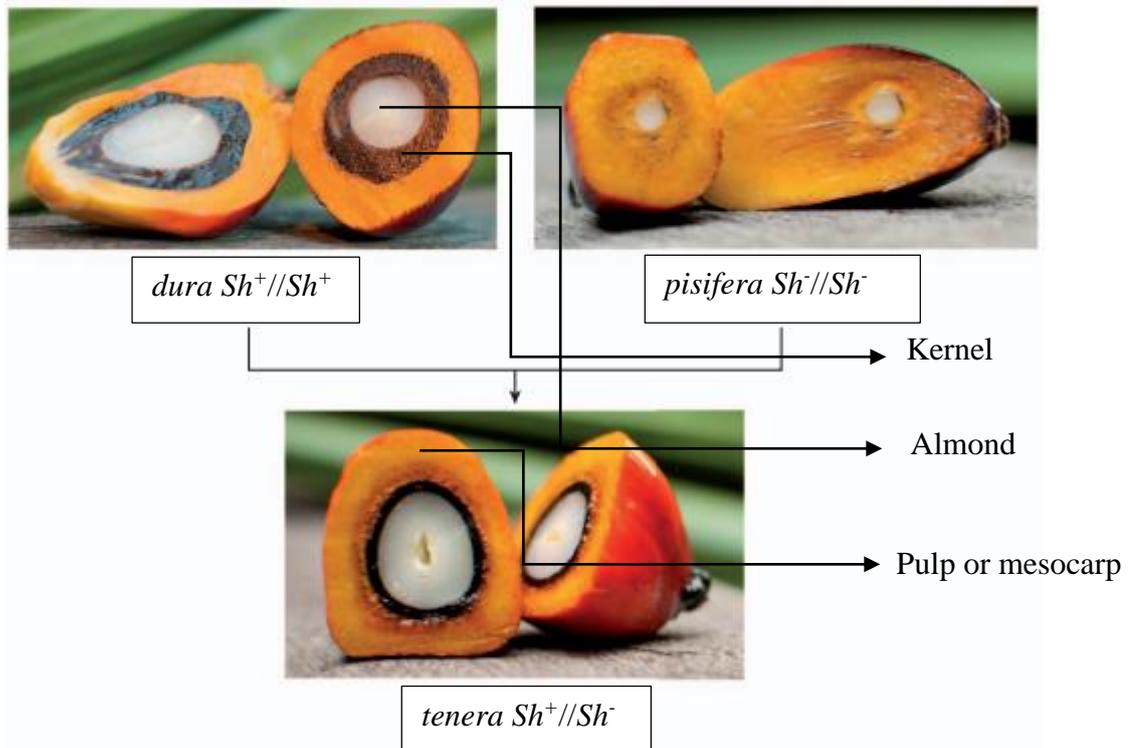


Fig. 9. Oil palm fruit forms (Singh et al., 2013a).

I.2.3.3. Fruit types

Pigmentation of fruits before maturity is at the origin of three fruits types: *Virescens*, *Nigrescens* and *Albescens* (Fig. 10).

Nigrescens is the most common and cultivated fruit type in commercial plantations. The apex of *Nigrescens* fruits is dark violet to black and the base is pale green to yellow, due to the presence of chlorophyll and anthocyanins in unripen fruits (Fig. 10a). As fruits grow, violet coloured area by anthocyanins increases while the green area reduces. At ripening, almost all the brown-coloured area turns to more or less deep red-orange due to the presence of carotenoids (Demol, 2002; Luyindula et al., 2005; Corley & Tinker, 2016).

Virescens plants have green fruits unripe and orange green at ripening with the top of the outer fruit remaining almost always greenish (Fig. 10b) (Demol, 2002). They are by far less common than *Nigrescens*, 0.5% in Nigeria, 0.7% in Angola and 6% in Cameroon (Rajanaidu, 1986; Hartley, 1988). *Virescens* trait is a qualitative trait controlled by a single dominant gene because homozygotes (*Vir//Vir*) and heterozygotes (*Vir//vir*) have shown identical phenotypes (Corley & Tinker, 2016). Palm oil quality of *Virescens* is of no economic interest (Demol, 2002).

Albescens type can be divided into two subtypes, *Albo-Nigrescens* (*Alb-Nig*) and *Albo-Virescens* (*Alb-Vir*). Before ripening, *Alb-Nig* fruits are black (Fig. 10c) and *Alb-Vir* fruits are

green light (Fig. 10d) so that they respectively become brown in the apex and yellow pale on the centre and the base, and yellow-green on the apex and light yellow on the centre and base, characteristic of a very low carotenoid content in the mesocarp. *Albescens* is the least common fruit type in natural palm groves (Demol, 2002; Luyindula et al., 2005).

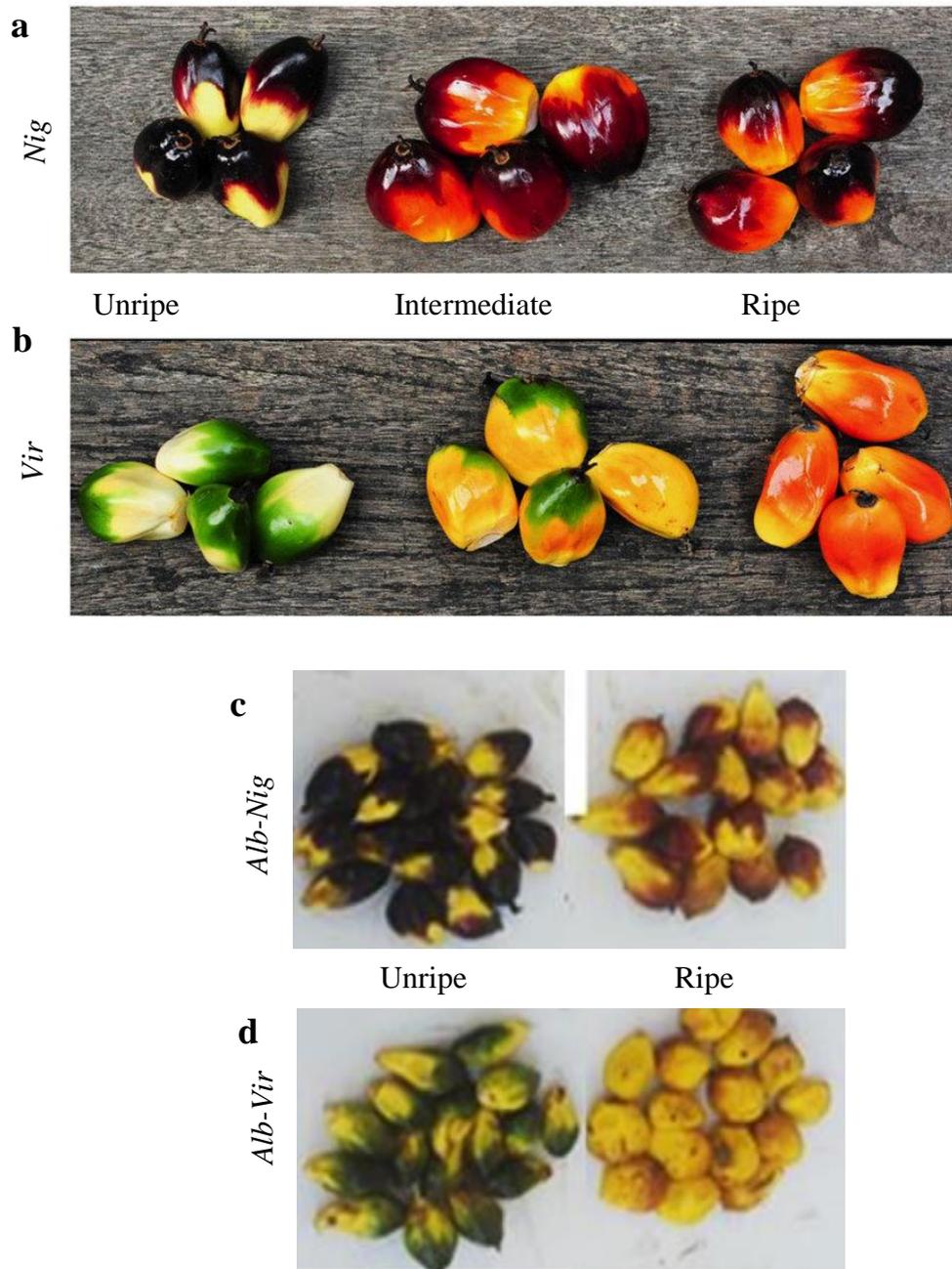


Fig. 10. Oil palm fruit types. a: fruits from *Nigrescens* (*Nig*) bunch, b: fruits from *Virescens* (*Vir*) bunch (Singh et al., 2014), fruits from *Albo-Nigrescens* (*Alb-Nig*) bunch and fruit from *Albo-Virescens* (*Alb-Vir*) bunch (Luyindula et al., 2005).

I.2.3.4. Mantled fruit type

Mantled fruit type was first named *poissoni* in 1918 after colonists' Poisson brothers settled in Cameroon. Diverse terms are used to designate these palms: palm trees with ears for

French, *diwakkawakka* for Germans and Dutch and mantled type for English. Mantled fruits contain up to six fleshy additional carpels derived from stamen primordia (Fig. 11) (Demol, 2002; Corley & Tinker, 2016). At first, mantled fruits seemed to present an interest because of a higher percentage of pulp on fruit and a fruit abscission delay. However, the oil content of the additional carpels is noticeably lower than that of the mesocarp itself. In addition, the number of fruits on mantled bunches is significantly lower than in ordinary types (Demol, 2002; Corley & Tinker, 2016) (Fig. 11).



Fig. 11. Transversal and longitudinal section of oil palm fruit. a: normal fruit, b: mantled fruits (Ong-Abdullah et al., 2015).

I.2.3.5. Reproduction system

Oil palm is a diploid naturally seed propagated plant and monoecious i.e., with male and female flowers carried on the same plant and usually in distinct inflorescences (set of flowers borne on spikelets), hence reducing selfing occurrences. The flowering of oil palm is continuous with an inflorescent bud in the axils of each leaf, constrained by external environment conditions and endogenous sexual cycles. Thus, inflorescent bud can develop into a male or female inflorescence which alternates during the individual plant lifetime (Jacquemard, 1995; Demol, 2002; Corley & Tinker, 2016). Consequently, oil palm is an obligate allogamous plant, with inflorescences enclosed in spathes tearing a few days before anthesis. A given palm tree produces barely two palms monthly (Jacquemard, 1995; Demol, 2002). Once the inflorescences have reached maturity, pollination will occur naturally thanks to pollinating agents (wind, insects, etc.) or for commercial seed production, under the control of a pollinating agent. Flower sex differentiation and inflorescence initiation start around 24 months before frond axils emerge. The nature of the future sex inflorescence is conditioned by the environment, thus

favourable environmental conditions induce female inflorescence production while hostile conditions favour male inflorescence production (Soh *et al.*, 2017).

I.2.3.5.1. Opened pollination

At maturity, male inflorescences produce pollen which emits a fragrant scent characteristic of anise, attracting insects making hence, open-pollination mainly entomophilic. In order to ease insect movements in both directions, male towards female inflorescences and conversely, papillae of different female flowers also emit a similar anise scent. Many species of insects are involved in oil palm pollination among which the predominant are weevils belonging to the genus *Elaeidobius*, with *Elaeidobius kamerunicus* from Cameroon being the main species (Syed, 1982; Corley & Tinker, 2016). Flower pollination can also be anemophilous but to a lesser extent (Syed, 1982).

I.2.3.5.2. Controlled pollination

Controlled pollination is used by seed producers to obtain the most yielding *dura* × *pisifera* progenies thanks to the best combinations of parents with known abilities. This laborious task is carried out following a rigorous and meticulous procedure to avoid any contaminations and obtain pure commercial seeds with the highest heterosis. Details about controlled pollination are described in (Rao & Kushairi, 1999; Periasamy *et al.*, 2002).

The first step consists on the identification of the target inflorescences, female or male, through weekly, then daily inspections. Once the inflorescence is identified, isolation just follows i.e., one week before the expected opening of the external spathe. Isolation of the female inflorescence involves spraying the flowers with formaldehyde, followed by bagging using a woven fiber bag with little pore size. Afterwards, impregnated cotton in an insecticide is then placed at the tied end of the bag to prevent penetration of any insects.

On the other side, male inflorescences are bagged following the same procedure as female inflorescences and are harvested at the anthesis, i.e., when inflorescences mature and fully open. Pollen collected from the inflorescence undergo a viability test and is used for controlled pollination afterwards.

I.2.3.6. Genetic resources for oil palm breeding

Genetic resources used for oil palm breeding in current research programs come from a narrow genetic base termed breeding populations of restricted origin (BPRO, (Rosenquist, 1986). Most of these palms come from Africa (La Mé, Yagambi, Ekona, etc.) and of plant sent from Africa and early planted in Asia thus forming decades after new geographical origins (Deli, AVROS).

The well-known *dura* Deli material used as a female parent in the commercial hybrid *tenera* seeds comes from four ancestors of an unknown area of Africa planted in Bogor Botanical Gardens in Java, Indonesia in 1848. All these four ancestors were phenotypically similar suggesting that they were from related palms or the same palm in Africa (Hartley, 1988). Progenies of these palms were first transferred in Sumatra plantations in Deli province in 1875 hence their name Deli, thenceforth planted and selected in other countries resulting in many other subpopulations bearing the names of their plantation localities. Indeed the Deli can be further divided into several subpopulations, such as Marihat Baris, Elmina, Ulu Remis, Dabou, etc. (Durand-Gasselin et al., 2000; Demol, 2002; Soh et al., 2003; Corley & Tinker, 2016).

Palms obtained in Eala Botanical Garden in Zaire, now Democratic Republic of Congo (DRC) from Djongo plant meaning the best in a local language during exchanges of breeding material were planted in 1923 in Sungai Pancur, Sumatra by *Algemeene Vereniging van Rubberplantera ter Oostkust van Sumatra* (AVROS), where its name comes from. As consequence, *pisifera* AVROS palms used as male parents in hybrid crosses are descendants of Djongo are characterized by their high oil yield, sturdy growth, thin shell, thick mesocarp, etc.

In Africa, there is an important genetic diversity currently used in breeding programs with the large majority being used as male parents (i.e., La Mé in Ivory Coast, Yangambi in DRC, Ekona in Cameroon, Calabar in Nigeria) in *tenera* hybrids and parents to a lesser extent as female (Angola). The La Mé population originated from 19 individuals selected from prospections made in the 1920s. The Yangambi population dated from the 1920s and originated from 10 to 20 *tenera*, included the Djongo palm which given its exceptional qualities, would have finally contributed more than 70% to the Yangambi population (Demol, 2002; Cochard, 2008; Corley & Tinker, 2016). The Ekona population originated from wild plantations located at Ekona, Cameroon that was further improved in the Unilever plantations.

I.2.3.7. Mass selection

Mass selection is the selection of individuals on the basis of their phenotypic performance. Therefore, its efficiency relies on the heritability of traits.

The genetic improvement of palm oil production started in the 1920s, in South-East Asia (Indonesia and Malaysia) and in what was then known as Belgian Congo (Demol, 2002; Corley & Tinker, 2016), and was based on mass selection.

In South-East Asia, the very narrow genetic base followed by several generations of selection led to the relatively homogenous and inbred breeding population Deli aforementioned (Demol, 2002; Corley & Tinker, 2016).

In Africa, as the source palms were of *dura*, *tenera* and *pisifera* types, the breeding approaches differed from those used in South-East Asia (Durand-Gasselín et al., 2000; Corley & Tinker, 2016). Breeding was less efficient in Africa, as it was complicated by the segregation of the fruit types in the crosses between the best *tenera* (Durand-Gasselín et al., 2000; Corley & Tinker, 2016). However, it led to the creation of the several breeding populations already mentioned (Demol, 2002; Cochard, 2008; Corley & Tinker, 2016).

Mass selection with the early breeding populations had been efficient as some components of oil yield had a moderate level of narrow-sense heritability h^2 such as PF (0.53) and BW (0.39) (Corley & Tinker, 2016). However, the other components (BN, FB and OP) had low h^2 (<0.25). This, and perhaps from knowledge of the advancement of breeding methodology from other crops, prompted the adoption of the more complex breeding schemes described below.

The breeding populations inherited from this period of mass selection can be classified into two complementary groups (A and B) based on the characteristics of their bunch production. Group A, mostly from South-East Asia (i.e., Deli population) and Angola, although the latter has been of lesser importance, produces a small number of big bunches. Group B, comprising the other African populations (with La Mé and Yangambi currently being the most widely used) and AVROS, produces a large number of small bunches (Meunier & Gascon, 1972). The complementarity of the FFB yield components traits in the two groups resulting in hybrid vigour explains the choice of A × B cross hybrid breeding approaches.

I.2.3.8. Current breeding schemes

The breeding schemes currently applied to improve oil palm yield involve two major improvements over mass selection: they exploit the hybrid vigour for bunch production that appeared in the A × B crosses, and they enable better estimates of genetic values. These schemes are mainly modified reciprocal recurrent selection (MRRS, Fig. 12), which generates sexual crosses, which account for the vast majority of oil palm commercial varieties grown in plantations; and clonal selection. They use mating designs, experimental designs and methods of statistical analysis that more efficiently separate the different genetic and environmental effects.

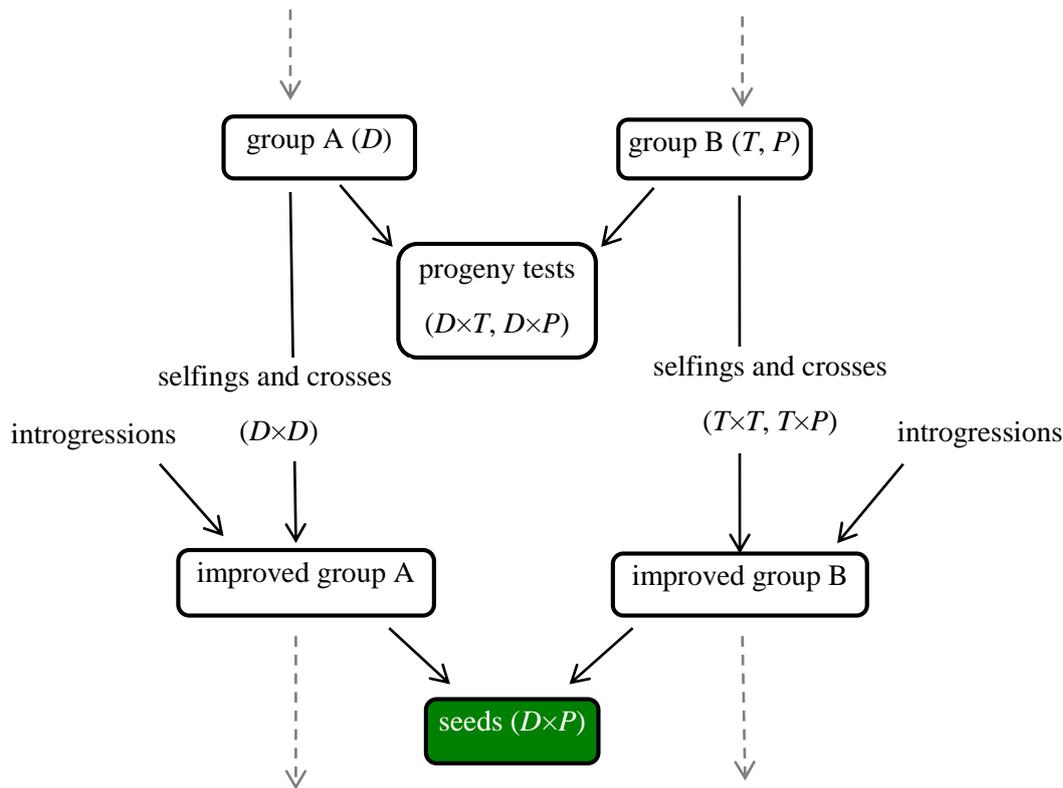


Fig. 12. Scheme of one cycle of modified reciprocal recurrent selection applied to oil palm (MRRS). *D*: *dura*, *T*: *tenera*, *P*: *pisifera*, green: commercial seeds (Nyouma et al., 2019).

I.2.3.8.1. Mating designs

In oil palm MRRS, the selected candidates are evaluated in hybrid crosses obtained according to NCM1 (NCM, North Carolina model) or NCM2 mating designs (Soh, 1999). The NCM1 is a hierarchical mating design in which each individual belonging to group B is crossed with a set of different individuals belonging to group A. If individuals in group A can be considered as genetically homogenous, NCM1 gives satisfactory estimates of the relative genetic or general combining ability values in group B. The NCM2 is a factorial design in which each B individual is crossed with the same set of A individuals (Corley & Tinker, 2016). This takes longer as several crosses have to be made per individual in group A, but is more suitable than NCM1 when genetic variability among the A individuals is not negligible or when the interactions between parents (i.e., specific combining abilities, SCA) need to be estimated.

I.2.3.8.2. Experimental designs

Once the crosses or the clones to be evaluated have been obtained, they are planted in field trials, usually according to randomized complete block designs (RCBD). The RCBD used in oil palm breeding usually has 10 to 50 families repeated three to six times in plots each of which contains 12 to 30 palms (Soh et al., 2017). Given the low planting density of oil palm

(normally 143 individuals per hectare), the trials require a large area (often >10 ha) whose environmental conditions are consequently subject to some heterogeneity. To better account for this heterogeneity, the complete blocks can be divided into incomplete blocks, i.e. comprising a sample of the evaluated families randomized within the complete blocks (Breure & Verdooren, 1995; Soh et al., 2017). Several experimental designs with incomplete blocks are thus commonly used for oil palm, including squared balanced or unbalanced lattices and alpha-plans (Soh et al., 2017). The results of evaluations of such trials using RCBDs and lattices have been published for hybrid crosses (Soh et al., 2017) and clones (Nouy et al., 2006). In experiments to study the genotype (G) × environment (E) interaction, the most commonly used design is the split-plot. In this case, E is the main treatment (planting density, fertilization, etc.) and G the sub-treatment (parents, hybrids or clones), which facilitates the management of the sub-plots and improves the statistical analysis, as the sub-treatment and the interaction effects are estimated more accurately (Soh et al., 2017). For instance, in a trial based on a split-plot design with planting density as the main treatment and hybrid crosses as sub-treatment, Rafii et al. (2013) found significant effects of G × planting density interactions on the average bunch weight.

I.2.3.8.3. Modified reciprocal recurrent selection

I.2.3.8.3.1. Principle

Reciprocal recurrent selection (RRS) was defined by Comstock et al. (1949) in maize. It relies on the joint and reciprocal improvement of two heterotic groups. A modified version of reciprocal recurrent selection (MRRS) was adapted for oil palm (Gascon & De Berchoux, 1964) and implemented by the *Institut de Recherches pour les Huiles et Oléagineux* (IRHO) in Ivory Coast (CNRA), Cameroon (IRAD), Benin (CRAPP) and Indonesia (SOCFINDO, IOPRI) (Meunier & Gascon, 1972; Corley & Tinker, 2016; Cochard et al., 2018). In oil palm, MRRS is justified by the fact that in A × B crosses the production of bunches is > 25% higher than in the parental populations (Gascon & De Berchoux, 1964). This is the result of the negative correlation between ABW and BN within each group, and from the complementarity of groups A and B for these two traits (Table III). Today, MRRS is used in many countries and, although its implementation varies among research centres, it generally follows the scheme described above (Fig. 12). However, a number of programs in Malaysia, Indonesia, and Papua New Guinea also practice the modified recurrent selection (MRS) or FIPS (family and individual palm selection) in which *dura* and *tenera* parents for further breeding are recurrently mass

selected and the *dura* × *pisifera* progeny testing is done to identify the parents, especially the *pisifera*, used for *dura* × *pisifera* seed production (Soh et al., 2017).

Table III. Origin of heterosis in oil palm for bunch yield.

	Annual number of bunches	Average bunch weight (kg)	Bunch yield (kg/an)
Group A	10	20	200
Group B	20	10	200
A × B hybrid	15	15	225

One cycle of oil palm MRRS (Fig. 12) starts with the selection of candidates from groups A and B and, after evaluation in hybrid progeny tests, the best ones will be selected among them. These candidates will then be used to produce the next generation, which will be used to produce seeds of *tenera* hybrids and to start a new MRRS cycle (Meunier & Gascon, 1972). In more detail, a cycle starts with phenotypic preselection prior to progeny tests. In group A, the individuals are selected based on their own phenotypic value for the traits with the highest heritability (mostly PF) and on the mean performance of their family (i.e., FIPS). In group B, the female sterility of *pisifera* means they can only be selected based on the mean value of their *tenera* full-sibs. For the same reason, and to be able to produce the following B generation, *tenera* individuals are also chosen by FIPS. Second, the combining ability of these individuals in hybrid crosses is evaluated in progeny tests, for the selection of low heritability traits and to finalize the selection of the traits subjected to the first stage of selection. For this purpose, the hybrids crosses are made according to the previously described mating designs, B individuals being crossed with three to five *dura* belonging to group A (Soh et al., 2010). These crosses are then evaluated in field trials, during which data are usually recorded from the third year after planting (i.e., at the beginning of production) to the tenth year. A long time is therefore required to obtain the genetic value of the progeny-tested individuals, resulting in long selection cycles lasting around 20 years. The resources required to carry out such long-term evaluations limit the number of individuals that are progeny tested, which results in the erosion of genetic diversity. To address this problem, new germplasms, for example originating from other breeding programs, are introduced (Jacquemard et al., 1997).

When analysing the phenotypic data of the progeny tests, the total genetic value of a hybrid cross is partitioned into the additive value or GCA of its parents or the non-additive or SCA of the cross. The GCA of a parent is the mean value of all the crosses that can be made between this parent and the parents of the other group, expressed as the difference from the

mean value of all possible hybrid crosses (Corley & Tinker, 2016; Gallais, 2011). The SCA of a cross is the difference between the observed value of the cross and the value predicted from the GCA of its parents (Gallais, 2011). It represents the interaction between its parents and usually results from dominance and/or epistatic effects (Stuber & Cockerham, 1966; De Souza, 1992). It can also result from the multiplicative interaction between two negatively correlated traits as BN and ABW for FFB production in oil palm. In this case, SCA may be present even in the absence of non-additive genetic effects (Schnell & Cockerham, 1992; Gallais, 2011). Finally, the parents with the best GCAs and/or resulting in the crosses with the best SCAs are selected. However, the SCAs for the components of oil palm yield are a much smaller source of variation among the hybrid performances than the GCAs, and are estimated with a lower accuracy than the GCAs (Cros, 2014). For these reasons, the selection is mostly made on the GCAs (Breure & Verdooren, 1995; Cros, 2014).

I.2.3.8.3.2. Statistical methods to estimate genetic values

According to the number of published articles, ANOVA is still the most widely used method to estimate GCAs in oil palm, and even to estimate the total genetic value of hybrid crosses without partitioning it into GCAs and SCAs (Breure & Bos, 1992; Okwuagwu *et al.*, 2008; Okoye *et al.*, 2009; Junaidah *et al.*, 2011; Noh *et al.*, 2012; Arolu *et al.*, 2016). To estimate the parental GCAs using ANOVA in a hybrid trial set up according to a RCBD, it can be considered that the yield y_{ijk} of cross $A_i \times B_j$ measured in block k is given by the model: $y_{ijk} = \mu + b_k + GCA_i + GCA_j + \varepsilon_{ijk}$, where μ is the phenotypic mean of the trial, b_k the effect of block k , GCA_i and GCA_j the parental GCAs and ε_{ijk} the error associated with the k^{th} replicate of the cross (Breure & Verdooren, 1995), with $y_{ijk} \sim N(E(y_{ijk}), \sigma^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$. The solutions of the model (i.e., the least square means), and in particular the parental GCAs, are obtained by the ordinary least squares' method. The SCAs are then obtained by subtracting the cross values expected from the parental GCAs from the mean cross values observed in the trial. ANOVA is useful for complete or balanced experimental designs and mating designs.

However, it is also possible to estimate the genetic values with the BLUP method, which is the standard approach for analyzing linear mixed models. BLUP was developed several decades ago to analyze highly unbalanced datasets in cattle breeding. Today it is widely used to estimate genetic effects in animals (Mrode, 2005) and in plants (Piepho *et al.*, 2008). BLUP has the following advantages (Soh, 1999): it is useful in analyzing unbalanced mating designs or experimental designs; and it makes it possible to consider a large number of trials at the same time, even without control families, and to account for covariances when modeling, for

example, the relationships among individuals, competition effects or spatial heterogeneity. Surprisingly, in oil palm it has only been used to estimate genetic values for yield components by a very limited number of research groups (Soh, 1994; Purba *et al.*, 2001; Cros *et al.*, 2015b). However, oil palm progeny tests are often carried out with complex and unbalanced designs, with a varying number of crosses per parent, crosses evaluated in several trials planted in different years, varying numbers of replicates and individual palms per cross, etc. The mating design is also sometimes not connected, i.e. that within a parental group, some parents are not connected (directly or indirectly) to the others by the same partners that belong to the other group, even though this can bias or make the GCA of some parents impossible to estimate (Breure & Verdooren, 1995; Soh *et al.*, 2017). Several studies have also shown that, in such complex situations, ANOVA was less efficient than BLUP in estimating the variances and/or the effects in the model (White & Hodge, 1989; Carvalho *et al.*, 2008; Piepho *et al.*, 2008; Hu, 2015). In addition, the pedigree of the oil palm breeding populations over several generations is generally known (Cros *et al.*, 2014; Corley & Tinker, 2016), and the relationships among selection candidates is useful information that can be included in the linear mixed model in order to more accurately estimate the genetic parameters and the genetic values.

In the case of hybrid crosses between two parental populations A and B, the linear mixed model used to estimate the parental GCAs and the cross SCA is:

$$y = X\beta + Z_1u_A + Z_2u_B + Z_3u_{AB} + \varepsilon$$

with: y the vector of observed phenotypes, β the vector of fixed effects, $u_A \sim N(0, 0.5A_A\sigma_{a_A}^2)$ and $u_B \sim N(0, 0.5A_B\sigma_{a_B}^2)$ the vectors of the GCAs of parents of groups A and B (random effects), respectively, and $u_{AB} \sim N(0, 0.25D_{AB}\sigma_{a_{SCAB}}^2)$ the vector of cross SCA, corresponding here to the dominance effects (random). X , Z_1 , Z_2 and Z_3 are, respectively, the incidence matrices associated to β , u_A , u_B and u_{AB} . $\varepsilon \sim N(0, I\sigma_\varepsilon^2)$ is the vector of residual effects and I is the identity matrix (in this example, residuals are assumed to be independent). $0.5A_A\sigma_{a_A}^2$, $0.5A_B\sigma_{a_B}^2$ and $0.25D_{AB}\sigma_{a_{SCAB}}^2$ are the variance-covariance matrices associated with u_A , u_B and u_{AB} , respectively. A_A and A_B are the matrices containing the values of additive relationships calculated with the pedigree of the A and B individuals, respectively, and D_{AB} is the matrix of dominance relationships between the crosses, and is obtained by the Kronecker product between A_A and A_B . $\sigma_{a_A}^2$ and $\sigma_{a_B}^2$ are the additive genetic variances of groups A and B, respectively, and $\sigma_{a_{SCAB}}^2$ is the dominance genetic variance of the crosses. The BLUP approach starts with estimation of the variances $\sigma_{a_A}^2$, $\sigma_{a_B}^2$, $\sigma_{a_{SCAB}}^2$ and σ_ε^2 . The most widely used method

for this purpose is restricted maximum likelihood (REML) (Xavier et al., 2016). Various algorithms have been developed to estimate the variance components with REML. The two main ones are the expectation-maximization algorithm (EM), which relies on the iterative updating of the residuals, variances and regression coefficients of fixed and random effects (Dempster et al., 1977); and the average-information algorithm, which relies on the creation of a gradient based on the mean of the expected and observed information (Gilmour et al., 1995). Second, the variances are used in the mixed model equations of Henderson, which give the model solutions, i.e. the vectors \hat{u}_A , \hat{u}_B and \hat{u}_{AB} for the genetic effects and the vector $\hat{\beta}$ for the fixed effects (Covarrubias-Pazarán, 2016). The solutions are named best linear unbiased estimators (BLUE), or solutions of the generalized least squares, for the fixed effects, and best linear unbiased predictors (BLUP) for the random effects (Mrode, 2005). The method also makes it possible to estimate the accuracy of the BLUPs, i.e., their correlation with the true genetic values that the model estimates. The accuracies are given by a theoretical formula using the diagonal of the variance-covariance matrix of the random effect considered and the prediction variance errors (PEV) associated with the BLUPs, which are easily obtained from the analysis. Thus, with the model presented here, the accuracy $r_{u_{A_i}, \hat{u}_{A_i}}$ of the GCA \hat{u}_{A_i} of parent A_i is:

$$r_{u_{A_i}, \hat{u}_{A_i}} = \sqrt{1 - \frac{\text{PEV}_{u_{A_i}}}{0.5(1+F_{A_i})\sigma_{a_A}^2}}, \text{ with } 0.5(1+F_{A_i})\sigma_{a_A}^2 \text{ the } i^{\text{th}} \text{ element of the diagonal of the}$$

variance-covariance matrix of u_A , and F_{A_i} the inbreeding coefficient of A_i (Cros, 2014). The application of this formula in oil palm showed that for the yield components, the hybrid progeny tests gave highly accurate GCAs, reaching on average 0.87 in group A and 0.91 in group B (Cros, 2014).

To promote the adoption of this method by the largest number of geneticists, in particular in the oil palm breeding community, in appendix 1, we provide a practical example of the estimation of the BLUP value of parents of oil palm hybrids using R software (R Core Team, 2017).

I.2.3.8.4. Clonal selection

The main use of clonal selection in oil palm is cloning the best *tenera* hybrid individuals. For this purpose, the *tenera* with the best phenotypes are chosen within the best crosses available in the MRRS program and are evaluated in clonal trials (Corley & Tinker, 2016). The interest of this method is based on oil palm heterozygosity, which generates genetic variability within the hybrid crosses, allowing selection of the best *tenera* individuals to be used as ortets

(source plants for cloning). The clones have the potential to further increase oil palm yield by 20% to 30% compared to sexual crosses (Corley & Law, 1997), and increases in yield of 13% (Nouy *et al.*, 2006) and 18% (Soh *et al.* 2003) have been empirically observed. One difficulty in clonal selection is to accurately estimate the genetic value of the hybrid individuals from their own phenotypic records, given the micro-environmental effects that are hard to control and are confounded with individual genetic values. This accuracy can be measured by the broad-sense heritability H^2 computed at the individual level. Soh *et al.* (2003), Nouy *et al.* (2006) and Potier *et al.* (2006) showed that H^2 ranged from 0 to 0.84 among yield components. In these conditions, it is possible to select ortets based on their phenotype for some traits, such as OP, but not for all yield components. Clonal field trials are thus required to finalize the evaluation of the ortets selected based on the traits with the highest H^2 . These trials allow a highly reliable selection of ortets, but lengthen the selection process by at least 10 years, corresponding to the time required to produce the clones from explants and to carry out the trial, thus allowing improved hybrids to catch up and reduce the advantage of clones.

Oil palm cloning has been slowed down by the appearance of abnormal floral morphogenesis in the field. The abnormal ramets, or mantled variants, produce abnormal flowers and fruits and bunch failure, leading to sterile palms (Soh *et al.* 2017). The epigenetic molecular mechanism that causes this abnormality was recently elucidated. The mantled variants were shown to result from hypomethylation during tissue culture of the Karma retrotransposon, located in the intron of the *DEFICIENS* gene. This altered its splicing and made it produce an additional transcript associated with the mantled phenotype (Ong-Abdullah *et al.* 2015; Soh *et al.* 2017). The understanding of this mechanism opens the way for the development of a molecular kit that will allow the early detection and elimination of abnormal ramets, thus boosting interest in oil palm cloning. Research is also underway to broaden the range of genotypes in which tissue culture is efficient (Soh *et al.* 2017). In addition, cloning opens the way for the production of genetically engineered palms. Indeed, tissue culture is an appropriate way to regenerate genetically modified tissue, and several genetic transformation methods have been successfully applied in oil palm (biolistic, transformation with *Agrobacterium* and microinjection) (Masani *et al.*, 2018).

1.2.3.8.5. Advantages and drawbacks

The current breeding schemes have the advantage of accurately estimating the genetic values, thereby enabling efficient selection, which, in turn, has enabled the significant genetic progress achieved so far. However, the schemes also have two drawbacks resulting from the difficulties involved in phenotyping. First, as mentioned above, the breeding cycle to produce

a new variety is long, around 20 years, whereas oil palm reaches sexual maturity relatively quickly (at three or four-year-old). The length of the cycle is mostly due to the phase of evaluation in progeny tests, as a long time is required to make the crosses, obtain the plants and above all, to carry out the field trial. Second, these schemes have low selection intensity, with - for example - fewer than 200 selection candidates progeny tested per population and cycle. The first stage of selection before the field trials (progeny tests or clonal trials) based on the phenotypic values for the most heritable traits seems to compensate for the reduced number of parents or clones evaluated, but this is not optimal. Indeed, the first stage of selection is made on a small number of traits and its accuracy is lower than selection based on progeny tests or clonal trials. Consequently, the individuals that would be the best considering their genetic value over all the yield components may be discarded before the field trials because they do not have the best phenotypic value for the trait or the few traits used in the first stage of selection. This even led to questioning the relevance of the first selection stage prior to field trials. For clonal selection, the possibility of randomly choosing the ortets before evaluating them in clonal trials has thus been considered by several authors (Corley & Tinker, 2016). However, to be efficient, this method would require exploring a large part of the genetic variability of the hybrid crosses where the ortets would be chosen, i.e., evaluating a large number of candidate ortets in clonal trials, which is not feasible in practice. New methods are therefore required to optimize the current breeding schemes.

I.2.4. Genomic selection

The first saturated genetic maps were produced at the end of the 1980s. They made it possible to detect QTLs (quantitative trait loci), leading to the idea of MAS. MAS has the potential to increase selection intensity and shorten the breeding cycles (Muranty *et al.*, 2014). Many QTLs related to oil palm yield have been identified (see for example Billotte *et al.* (2010), Pootakham *et al.* (2015), Tisné *et al.* (2015), Ting *et al.* (2018)). However, for complex traits such as yield that are under the control of a large number of genes with small effects, the efficiency of the approach is limited, in particular in the case of small population size (Muranty *et al.*, 2014), because it overestimates the effect of the strong QTLs and fails to exploit weak QTLs, as their effect does not appear to be significant (Muranty *et al.*, 2014). A more efficient approach, genomic selection (GS), was consequently developed (Meuwissen *et al.*, 2001). Its practical implementation was made possible by progress in genomics, in particular in next generation sequencing (NGS) and high throughput genotyping. Today, GS is used in animal breeding, particularly in dairy cattle, where it has doubled the rate of the genetic progress

(Wiggans et al., 2017). In plants, it is progressively being incorporated in breeding schemes, and it is expected to significantly increase their efficiency (Varshney et al., 2017).

In oil palm, the use of GS to select the parents of the hybrid crosses for yield traits has already been investigated in several studies. They evaluated its ability to reduce the length of the breeding cycles, by avoiding field trials in some cycles, and to increase selection intensity, by the application of selection to a larger number of candidates than with the current method (Fig. 13). The results are promising and are detailed below. So far, no study has been published regarding the use of GS to select ortets, but its potential is likely also high, as suggested by the positive results obtained in other species, and in particular in other perennial tropical crops like eucalyptus (Durán et al., 2017) and rubber tree (Cros et al., 2019).

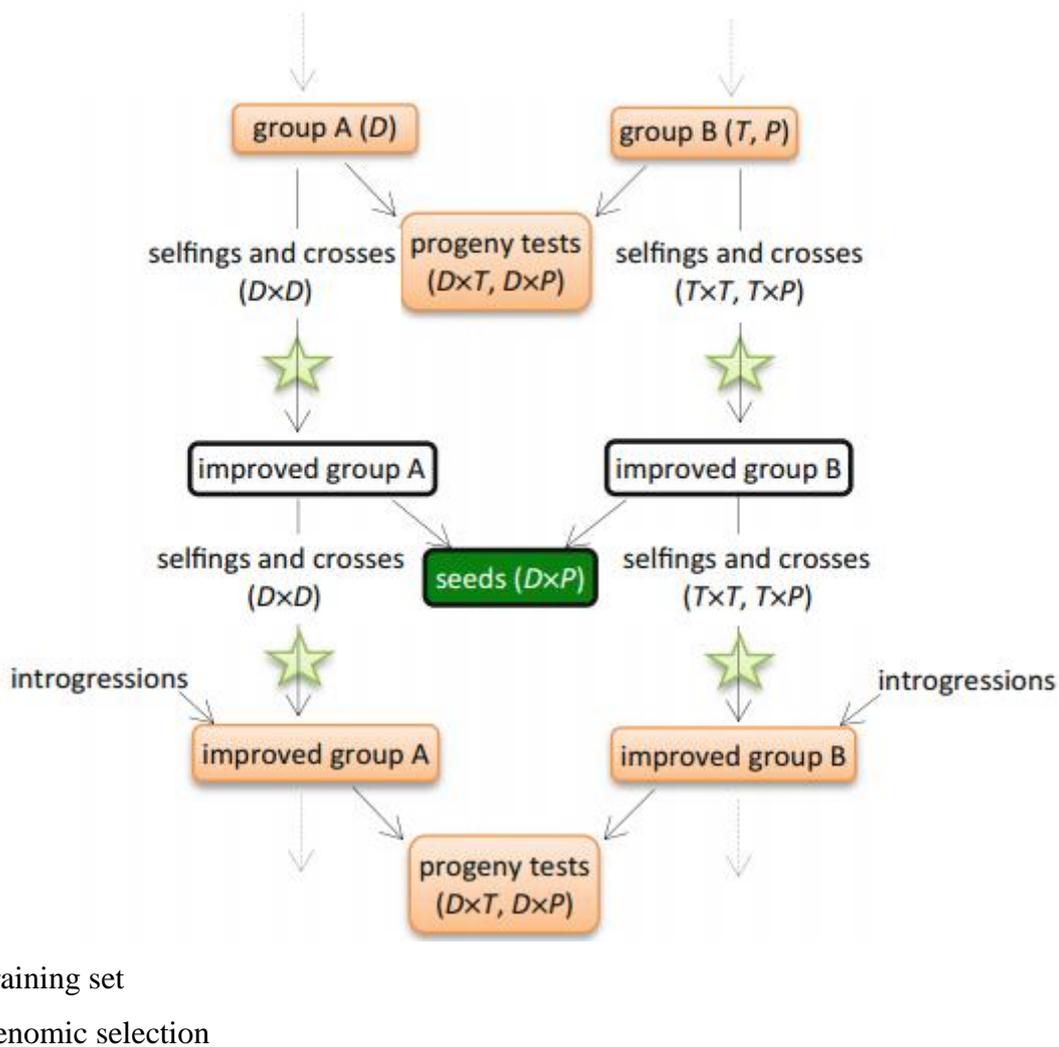


Fig. 13. Possible scheme of genomic modified reciprocal recurrent selection applied in large populations of seedlings to increase selection intensity (cycles 1 and 2) and shorten breeding cycles (cycle 2) of oil palm. *D*: *dura*, *T*: *tenera*, *P*: *pisifera*, green: commercial seeds (Nyouma et al., 2019).

I.2.4.1. Principle

GS is MAS for quantitative traits using high-density molecular markers covering the whole genome, in order to have every QTL in linkage disequilibrium with at least one marker. What mainly differentiates it from QTL-based MAS is the joint exploitation of strong QTLs (i.e., whose effect would be shown to be significant in a QTL analysis) and of weak QTLs (not significant). Its goal is to predict the genetic value of selection candidates, usually with no data on their performance (i.e., depending on the breeding situation concerned, with no known phenotype or no progeny tests). For this purpose, GS uses the genotypic and phenotypic data of a population called the training (or calibration) population and a linear mixed model that can predict the additive genetic value (GEBV, genomic estimated breeding values) or the total genetic value (i.e. including the non-additive effects) of the selection candidates (Heffner *et al.*, 2009) (Fig. 14). GS, therefore, has the potential to reduce phenotyping, thus making it possible to shorten the breeding cycle and/or to increase selection intensity.

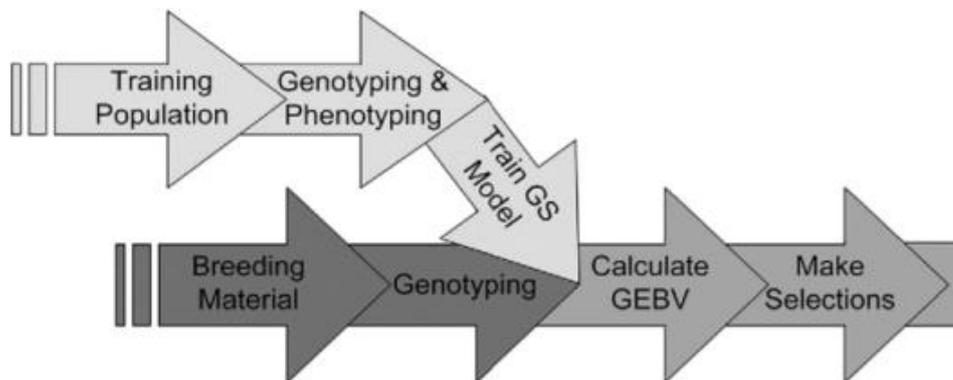


Fig. 14. Diagram of genomic selection (Heffner *et al.*, 2009).

The efficiency of GS is assessed by computing its selection accuracy (r_{GS}), i.e., the correlation between the genetic value estimated with the genomic model (GEGV) and the true genetic value (TGV) in a set of individuals used as the validation population. However, in empirical studies, the true genetic value is unknown, and the genetic value estimated with the genomic model is therefore correlated with an estimate of the true genetic value (EGV), obtained with the phenotypic data available on the validation individuals, i.e., their own phenotypic records or the phenotypes of their progenies. This correlation is named prediction accuracy. The difference between selection accuracy and prediction accuracy depends on the reliability of the EGV (Lorenz *et al.*, 2011). GS accuracy is crucial to evaluate the potential of GS as it is directly related to the rate of the genetic progress or rate of selection response $R = r_{GS} \times i \times \sigma_g/L$, with σ_g the genetic variance and L the generation interval (Falconer & Mackay, 1996). However, a comprehensive comparison of GS and conventional selection

requires considering their respective selection accuracy, selection intensity and generation interval. Indeed, even in a situation where GS accuracy would be lower than the accuracy of the conventional phenotypic evaluations, GS can still increase R if it allows a sufficient decrease in the generation interval and/or increase in selection intensity.

GS accuracy is affected by several parameters, including marker type and density, distribution of QTL effects, linkage disequilibrium between markers and QTLs, the size of the training population, the relationship between the training and selection populations, trait heritability and statistical methods of prediction (Lorenz *et al.*, 2011; Grattapaglia, 2014). In practice, GS accuracy is usually estimated by cross-validation at a single experimental site (Cros *et al.*, 2015b; Kwong *et al.*, 2017a,b) or by between-site validation (Cros *et al.*, 2017). However, single-site cross-validations may overestimate accuracy, and it is therefore preferable to have at least two sites to evaluate GS (Lorenz *et al.*, 2011).

I.2.4.2. Molecular data

GS generally uses single nucleotide polymorphism markers (SNPs). They are abundant on the whole genome, have a low mutation rate (Oraguzie *et al.*, 2007) and can easily be genotyped at a reasonable cost. In oil palm, given the molecular resources available at the time, the first empirical studies were made with microsatellites (SSR, simple sequence repeats) (Cros *et al.* 2015b; Marchal *et al.*, 2016). However, GS studies in this species now use SNPs from genotyping-by-sequencing (GBS) (Cros *et al.*, 2017) or SNP arrays (Kwong *et al.*, 2016, 2017a,b; Ithnin *et al.*, 2017). This allowed reaching higher densities, which contributed to achieve higher accuracies. Thus, Kwong *et al.*, (2017b) using 135 SSRs obtained mean GS prediction accuracies of 0.21 over palm oil yield components, against 0.31 with 200K SNPs.

GS accuracy normally increases with the number of markers until it reaches a plateau (De Los Campos *et al.*, 2013; Cros, 2014). In oil palm, the effect of marker density on the GS accuracy for yield components has been evaluated in three studies. When predicting the performance of unevaluated hybrids, GS accuracy started plateauing with 500 and 2,000 SNPs in Cros *et al.* (2017) and between 200 and 400 SNPs in Kwong *et al.* (2017a), depending on the trait. The two studies did not consider the same populations, but the smaller number of SNPs required in Kwong *et al.* (2017a) likely resulted from the fact that the SNPs were chosen based on the association scores estimated in a genome-wide association study, and not randomly, as in Cros *et al.* (2017). When predicting the GCA of progeny-tested individuals, Marchal *et al.* (2016) showed that GS accuracy plateaued with 160 SSRs in group A and 90 SSRs in group B. The marker density required to reach the maximum GS accuracy, therefore, varies depending

on the type of marker, the marker sampling method, the trait and the population. However, the marker density needed in oil palm is lower than is generally the case in other species due to the high rate of inbreeding in oil palm breeding populations, i.e. to their small effective size (Cros *et al.*, 2014).

Genotyping generates missing data. There are very few missing data with SNP arrays (< 1% in Kwong *et al.* (2016)) and SSRs (< 3% in Cros *et al.* (2015b)), but they can reach significant proportions with GBS (13.2% in Cros *et al.* (2017)). The GS statistical models cannot deal with missing molecular data, which therefore have to be imputed. This consists in replacing them by the most likely genotype. In practice, the imputation method is likely of no importance when the percentage of missing data is low. In this case, the missing data can be replaced by the genotype with the highest frequency for the marker considered in the population concerned, as in Kwong *et al.* (2017a). With more missing data, more sophisticated imputation approaches are recommended. Many methods are available for this purpose (Wang *et al.*, 2016). Currently, only the BEAGLE software (Browning & Browning, 2007) has been used to impute missing molecular data in GS studies on oil palm. Cros *et al.* (2017) showed that taking pedigree information into account for imputation made BEAGLE more efficient. However, they also noted that, for a given number of markers, using those with the lowest percentage of missing data resulted in higher GS accuracy than using random markers, which suggests that imputation could be improved.

I.2.4.3. Training and application populations

GS accuracy normally increases with the size of the training population (Lorenz *et al.*, 2011; Grattapaglia, 2014) and with the relationship between training and application individuals (Pszczola *et al.*, 2012). In oil palm, GS accuracy was observed empirically to be strongly affected by the relationship between training and application individuals (Cros *et al.*, 2015b), suggesting that the use of GS in full-sibs or progenies of the training individuals would maximize accuracy. To increase the size of the training set, it is possible to aggregate data from consecutive breeding cycles. Simulations in oil palm showed that using data from two cycles increased the per cycle response to selection by more than 10%, mainly as a result of higher selection accuracy (Cros *et al.*, 2018). Although this aggregation of data reduces the relationship between training and application populations, this is more than counterbalanced by the doubling of the training population.

Several strategies can be used to optimize the training and application populations. For instance, the CDmean criterion, derived from the generalized coefficient of determination, can

optimize the sampling of individuals that have to be phenotyped among a set of genotyped individuals, in order to form the training population (Rincent *et al.*, 2012). In oil palm, the CDmean proved to be efficient for GS as it maximizes its accuracy (Cros *et al.*, 2015b). However, further improvements are possible: for example, another optimization criterion recently developed to define training populations, CDpop, could be more efficient for oil palm as it is specific to highly structured populations (Rincent *et al.*, 2017).

I.2.4.4. Models and statistical methods for genomic predictions

Genomic predictions are made with frequentist and Bayesian statistical approaches (Varshney *et al.*, 2017). Some methods estimate an effect associated with each marker, while other methods give the genetic values directly without estimating marker effects. Genomic predictions exploit two types of information, the relationship between training and application populations, and the linkage disequilibrium between markers and QTLs (Varshney *et al.*, 2017).

In methods that estimate marker effects, the base (i.e., purely additive) genomic linear mixed model is of the form: $y = X\beta + Zm + e$, where y is the vector of data records ($n_{\text{ind}} \times 1$), β the vector of fixed effects (mean, trials, blocks, etc.) associated with incidence matrix X , m the vector containing the substitution effect of each SNP ($n_{\text{SNP}} \times 1$) with incidence matrix Z ($n_{\text{ind}} \times n_{\text{SNP}}$) containing the molecular data coded in the number of copies of the most frequent allele (0, 1 or 2), e the vector of residuals ($n_{\text{ind}} \times 1$), n_{ind} the number of individuals in the training population and n_{SNP} the number of SNPs (Soh *et al.* 2017). The effects m and e are random. The GEBV of selection candidate i is given by summing the SNP effects over the whole genome according to the formula: $\text{GEBV}_i = \sum_{j=1}^{n_{\text{SNP}}} Z_{ij} \hat{m}_j$, with \hat{m}_j the estimated effect of SNP j . Depending on the way the marker genetic variance (σ_m^2) is treated, two types of methods can be distinguished (Soh *et al.* 2017). First, some methods consider that marker effects are sampled according to a normal distribution with a variance common to all markers, which is relevant for traits following the infinitesimal model. This is the case of random regression BLUP (RR-BLUP) (Meuwissen *et al.*, 2001) and Bayesian random regression (BRR) (Pérez *et al.*, 2010). Second, as the genetic determinism of some quantitative traits may include loci with strong effects, other methods such as Bayes A, Bayes B (Meuwissen *et al.*, 2001), Bayes C π , Bayes D π (Habier *et al.*, 2011) and Bayesian LASSO (De Los Campos *et al.*, 2009) attribute marker specific genetic variances.

The most widely used method to estimate GEBV directly is the genomic best linear unbiased predictor (GBLUP). The basic difference between GBLUP and conventional BLUP presented above is the use of genomic (instead of genealogic) information to compute the

relationship matrix, called the \mathbf{G} matrix in GBLUP. The \mathbf{G} matrix has the advantage of accounting for the random sampling of alleles at meiosis (Mendelian sampling) and thus gives realized relationships, making it possible to obtain the GEBV of unevaluated individuals. Also, genomic data are not affected by pedigree errors in the families used in the breeding program. By contrast, the pedigree-based \mathbf{A} matrix gives expected relationships (Habier *et al.*, 2007; VanRaden, 2007), and therefore does not differentiate between individuals within families, cannot capture relationships that do not appear in the pedigree records and gives erroneous values in the case of illegitimacy. The base model used with GBLUP is: $y = \mathbf{X}\beta + g + e$, with g the vector ($n_{\text{ind}} \times 1$) of GEBVs following $N(0, \mathbf{G}\sigma_g^2)$, σ_g^2 the additive variance and \mathbf{G} ($n_{\text{ind}} \times n_{\text{ind}}$) the genomic relationships matrix. With SNP markers, the \mathbf{G} matrix is usually computed according to VanRaden (2007). GBLUP is equivalent to RR-BLUP under the assumption of normality of marker effects and has the advantage of being simple to implement with existing software and of having a reasonable computation time.

Various modeling approaches have been used for genomic predictions in oil palm. The base GS models described above were used in each parental group separately, with data records consisting of parental performances in crosses with the other group, i.e. GCAs (Cros *et al.*, 2015b) or testcross phenotypic means (Wong & Bernardo, 2008), and parent genotypes. Ithnin *et al.* (2017) and Kwong *et al.* (2017b) applied similar models but used parental phenotypes as data records. They obtained low to intermediate GS prediction accuracies but, as parental phenotypes may not reflect performance in hybrid crosses due to gene-frequency differences between parental populations and non-additive effects (Wei *et al.*, 1991; Baumung *et al.*, 1997; Vitezica *et al.*, 2016), the relevancy of such accuracies for hybrid breeding is questionable. Kwong *et al.* (2016) studied GS with a population consisting in a mixture of Deli, group B and hybrid individuals. They obtained a prediction accuracy of 0.65, which could have possibly been improved by the use of a model designed to jointly consider parental and hybrid data, like in Vitezica *et al.* (2016). Accuracy of GS could also be improved by a single-step GBLUP (ssGBLUP) which blends realized relationship of genotyped individuals with the genealogical relationship of non-genotyped individuals to calculate GEBV. This increases the size of the training set by taking into account ungenotyped individuals for which phenotypes are available. In oil palm, this could be used to include in the training set phenotyped individuals for which DNA can no longer be obtained, such as individuals evaluated in past progeny tests. In eucalyptus, using additional phenotypic information from non-genotyped individuals thus increased GS prediction accuracies by up to 75% (Cappa *et al.*, 2019). Other studies used the conventional MRRS model replacing genealogical relationship matrices by genomic matrices

to jointly predict the GEBV of A and B candidates (Cros *et al.*, 2015b, 2017, 2018; Marchal *et al.*, 2016). In order to increase the training size, this method was adapted to include molecular data of individual hybrids, taking into account the parental origin of marker alleles (Cros *et al.*, 2015a). This gave the highest selection accuracies for unevaluated parents, and thus proved to be more efficient than using only parental genotypes to train the model. Kwong *et al.* (2017a) also used molecular data of individual hybrids, but did not consider the parental origin of alleles. So far, the usefulness of modeling the parental origin of marker alleles in oil palm hybrids genotypes has not been investigated. Further studies thus remain necessary to identify the optimal prediction model, in particular depending on the nature of the training data.

In addition, a wide range of statistical methods has been applied to analyze these models, and comparisons showed that they did not significantly affect the accuracy of GS (Cros *et al.*, 2015b; Ithnin *et al.*, 2017; Kwong *et al.*, 2017b). This suggests that the components of palm oil yield are highly polygenic and follow the infinitesimal model.

I.2.4.5. Information captured by markers

Without optimizing the training and validation populations, prediction accuracies ranging from 0.14 and 0.73 were obtained for various yield components, confirming the ability of GS models to predict the genetic value of unevaluated selection candidates (Cros *et al.*, 2017; Kwong *et al.*, 2017a,b). In particular, for five yield components (FFB, OP, BN, BW and PF), the GS model predicted the performance of unevaluated hybrid crosses with higher accuracy than a control model using pedigree data instead of markers (Cros *et al.*, 2017). This showed the ability of GS to capture genetic differences within full-sib families (i.e., the Mendelian segregation term) in addition to genetic differences between families, enabling the selection of the best individuals within the best families, as currently done among the individuals that are progeny tested. The same conclusion was reached in Kwong *et al.* (2017b), where GS prediction accuracies above zero, ranging from 0.18 to 0.47, were obtained in a GS evaluation considering a single full-sib family. Similarly, Cros *et al.* (2015b) obtained GS prediction accuracies above 0.5 within full-sib families. However, the latter study also showed that GS could also, depending on trait and population, fail to capture Mendelian segregation. In this case, GS predictions only revealed, at the best, between-family differences.

I.2.5. Genetic progress

The first GS study in oil palm was a simulation study (Wong & Bernardo, 2008), starting with an initial breeding population derived from the selfing of a hybrid. Two cycles of conventional breeding were simulated. At each cycle, the breeding population was crossed with a tester to allow phenotypic selection for yield performance, and the selected individuals were

crossed to produce the new generation. With MAS (QTL-based MAS and GS), the initial population was also genotyped and used to estimate marker effects, and in the following cycles, phenotypic selection was replaced by selection on markers. This reduced the length of the breeding cycles and enabled three consecutive selection cycles on markers, with a total number of years over the four cycles equivalent to the two cycles in conventional phenotypic selection. The authors found that GS and conventional selection outperformed QTL-based MAS in terms of selection response, while GS outperformed conventional selection when the population size reached 50 to 70 individuals, and then increased selection response by 4% to 25%, depending on population size, heritability and number of QTLs.

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

These simulation results promise a revolution in the genetic improvement of oil palm yield. However, this needs to be put into perspective by the empirical studies that, even if they showed that GS accuracies could be high, also revealed that GS was not efficient for all yield components. Indeed, for some traits, the GS model did not predict the genetic value of

unevaluated individuals better than a control model using pedigree data instead of markers (Cros *et al.*, 2015b, 2017). Yet, the simulations showed that the main advantage of GS was its ability to shorten the breeding cycles by avoiding field evaluations in some cycles, and this is only possible if GS is efficient for all the yield components that are currently the subject of phenotypic selection. Otherwise, the progeny tests remain necessary in all breeding cycles. Therefore, the practical application currently envisaged to start implementing GS in oil palm is a two-stage scheme, with an initial stage of genomic selection prior to progeny tests. This would be better than the current first stage of phenotypic selection for two reasons. First, the number of yield components for which GS is efficient is greater than the number of traits currently subjected to phenotypic preselection. Second, the current selection prior to progeny tests is made on the parental phenotypes, even though, as already mentioned, they may be poor indicators of performance in hybrid crosses. By contrast, this would not be a problem for genomic predictions obtained with a model calibrated on hybrid phenotypes. The potential of genomic preselection was quantified based on the GS accuracies empirically obtained by between-site validation for bunch production, a trait which is normally not subjected to phenotypic selection prior to progeny tests in the current schemes (Cros *et al.*, 2017), and the study showed that this would increase the performance of the selected hybrids by more than 10% compared to a method without preselection, thanks to higher selection intensity.

To be applied in practice, GS must also result in annual genetic progress per unit cost higher than current selection methods. Although GS generates additional costs related to genotyping, these costs are low in comparison to the cost of phenotyping. Thus, Jacob *et al.* (2017) indicated that, even assuming a genotyping cost per sample as high as 300€, which seems to be the maximum possible price for a 300K SNP array, the ratio of genotyping/phenotyping costs lays below 1/20. In addition, these extra costs could possibly be offset by a reduction in phenotyping costs, when it is possible to manage without some field evaluations. In this case, Wong & Bernardo (2008) found that with a genotyping cost of US\$0.15 per datapoint, corresponding to genotyping prices for SNPs, the cost per genetic progress unit was 35% to 65% lower with GS than with conventional selection.

CHAPTER II. MATERIAL AND METHODS

II.1. Material

II.1.1. Study sites and experimental designs

The current study has been carried out in two sites, at Aek Loba Timur (ALT) at 2° 39' North – 99° 42' East and Aek Kwasan division VI (AK) at 2° 38' North – 99° 37' East both located in North Sumatra (Fig. 15), on the SOCFINDO estate (Indonesia) and with 9 km of distance separates them. They are both situated at around 50 km from the sea level on deep loamy sand soils, with low water deficit and high insolation, and benefiting from standard cultural practices and the same protocol for data record (Potier *et al.*, 2006; Cros *et al.*, 2017).

The experimental designs used in both sites were either balanced lattice of four to five ranks or randomized complete block designs (RCBD). ALT is constituted of 28 trials (Fig. 16) and AK is divided into, AK1 composed of seven trials and AK2 composed of 19 trials (Potier *et al.*, 2006).

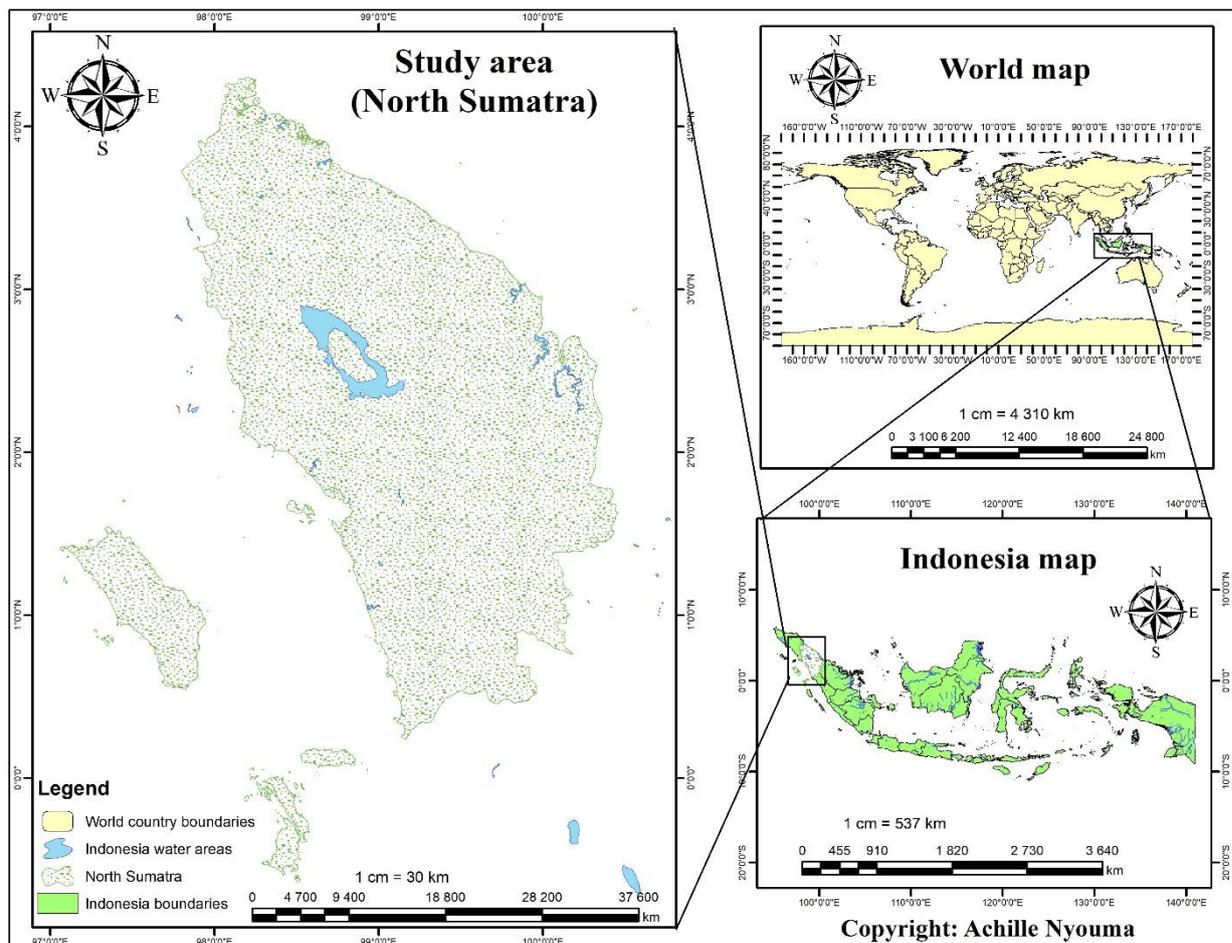


Fig. 15. Map of the study area.

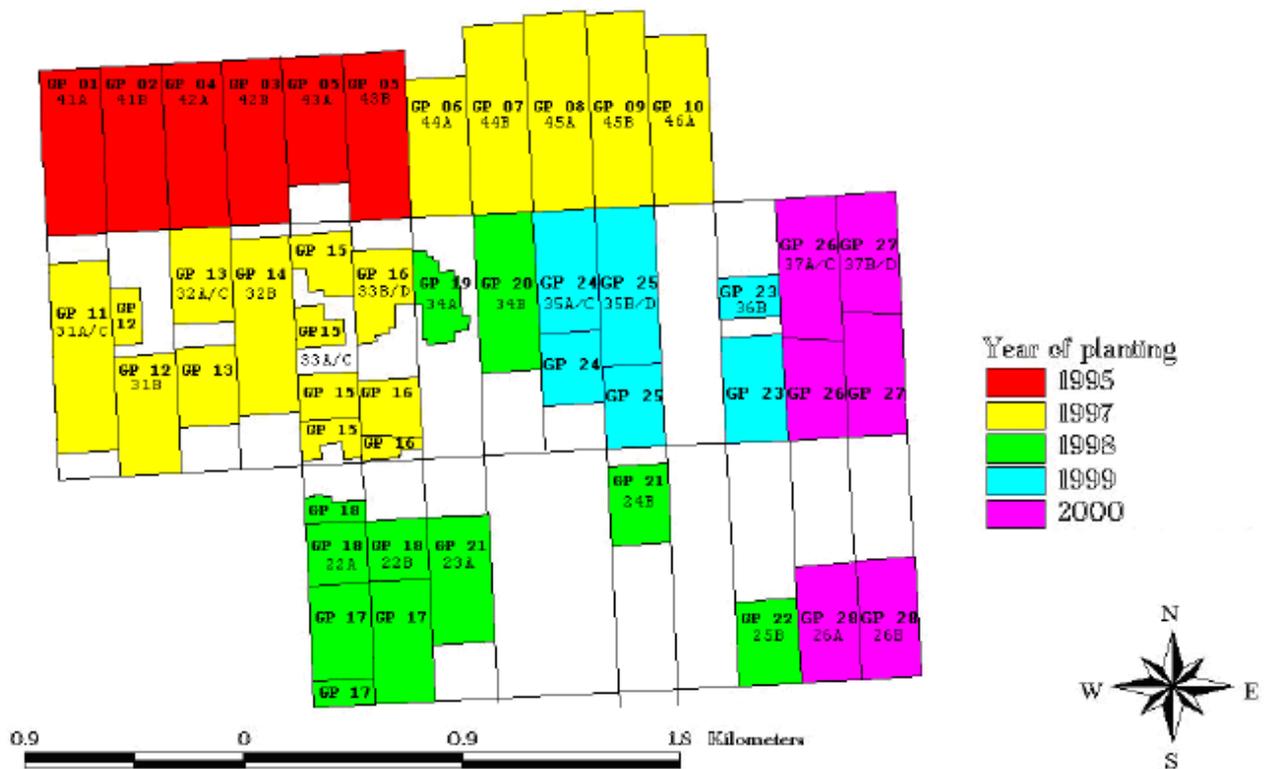


Fig. 16. Location plan of the 28 trials (GP) of Aek Loba Timur (ALT).

II.1.2. Plant material

To evaluate the efficiency of genomic selection (GS) for clonal selection, the plant material used to train the GS models comes from controlled crosses between Deli and La Mé (LM) individuals. For bunch production predictions, the training set was composed of 295 progeny-test crosses planted from 1995 to 2000 at 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 IV). The pedigrees of these populations are known over several generations.

The validation set was composed of 42 Deli \times 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 IV). The 42 ortets were chosen among individuals from various hybrid crosses planted on seven trials of an earlier set of progeny tests, located at AK1. 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.

Table IV. Characteristics of the datasets used for training and validation for clones.

	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

To evaluate the effect of the genotyping strategy to optimize prediction accuracy, the parental populations used comprised two groups, group A, mostly consisting of an Asian population (Deli) and, to a lesser extent, Angola, and group B, composed of other African populations (La Mé from Ivory Coast, Yangambi and Lisombe Kinshasa from the Democratic Republic of the Congo, Nifor from Nigeria and Sibiti from the Republic of Congo). Nine yield components were assessed: three bunch production traits BN, FFB, ABW, and six bunch quality traits i.e., AFW, NF, FB, PF and OP, and OER.

The training set for bunch production contained 352 A×B *tenera* hybrid crosses including 123 parents in group A and 121 parents in group B for a total of 22,656 hybrid individuals. Among these crosses, only 341 could be used as a training population for bunch quality traits, because phenotypic data for these traits was only available for a few crosses; the crosses involved 121 parents in group A and 118 parents in group B, for a total of 14,985 hybrid individuals (Table V). Training palms were planted from 1995 to 2000 in ALT.

For the validation of bunch production, we used a set of 213 A×B *tenera* hybrid crosses involving 71 parents in group A and 49 parents in group B, with a total of 13,399 hybrid individuals for bunch quality and 10,339 for bunch production. Palms destined for the validation set were planted between 2005 and 2009 in 19 trials in the AK2 (Table V).

Table V. Composition of the datasets used for training and validation for hybrids.

	Training population		Validation population	
	bunch production	bunch quality	bunch production	bunch quality
Number of crosses	352	341	213	213
Number of individuals (genotyped)	22,656	14,985	13,399	10,339
Average number of individuals per cross (min–max)	64 (17-503)	44 (21-292)	63 (25-680)	48 (19-493)
Number of group A parents	123	121	71	71
Number of group B parents	121	118	49	49
Age at time of data collection (years)	3-7	5-9	3-7	5-9

II.2. Methods

II.2.1. Evaluation of the efficiency of genomic selection for clonal selection

II.2.1.1. Phenotyping

All the individuals, i.e., the training hybrid crosses, the 42 hybrid ortets and their ramets, were phenotyped for eight traits. Five traits were assessed for bunch quality: 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); and three traits for bunch production: bunch number (BN), average bunch weight (ABW), and total bunch production (FFB). For quality traits, data were collected when plants were from five to nine years old at ALT and from six to nine years old at AK1. For production traits, data were collected when the plants were from three to seven years old in both sites.

II.2.1.2. Genotyping

Molecular data were obtained by GBS (Elshire *et al.*, 2011; He *et al.*, 2014) for the 42 ortets, 93 Deli and 91 La Mé parents of the training hybrid crosses (Table IV). Ortets genotypes were obtained from two or three samples collected on different ramets (thus allowing

controlling the legitimacy of the ramets). DNA extraction and GBS were performed as described in Cros *et al.* (2017), using the *Pst*I and *Hha*I restriction enzymes. The raw fastq sequence data were processed with Tassel GBS v. 5.2.44 (Glaubitz *et al.*, 2014), using the Bowtie2 software for alignment (Langmead & Salzberg, 2012), and VCFtools 0.1.14 (Danecek *et al.*, 2011). The indels were discarded, the datapoints with depth below five were set to missing, the SNPs that were not biallelic, with more than 75% of missing data or on the unassembled part of the genome were discarded. This resulted in a dense genome covering, with 15,054 SNPs. The average percentage of missing data was 23.08% (3.64% - 43.42% per individual). To explain the 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.

II.2.1.3. Imputation of missing SNP data and phasing

Imputation of missing SNP data and phasing were carried out with Beagle 4.0 (Browning & Browning, 2007). 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. 17. 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, for the missing parents that were obtained through selfing, the selfed grandparents were used in the pedigree instead of the actual parents, as grandparental DNA was available (for the other steps of the analysis that required a pedigree, the real pedigree was used). As some ortets remained with one parent that was not genotyped and that did not originate from selfing, we used a home-made R script to recover the parental origin of ortet phases. For each ortet, this script considered the two phases, one after another, and checked all along the genome if similar blocks of consecutive SNPs were found in the Deli and La Mé parent. Each ortet phase was finally assigned to the parental population with the highest number of SNP blocks specific to the population that was found on the considered ortet phase.

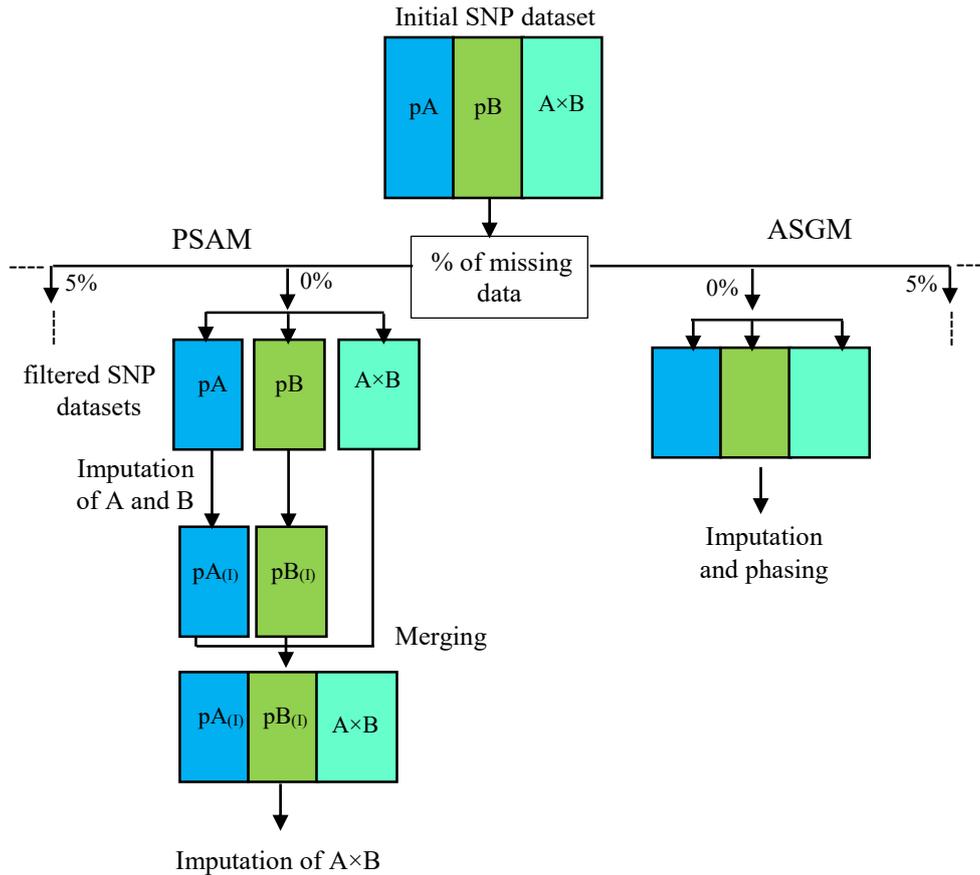


Fig. 17. Imputation and phasing scheme for the production of the SNP datasets used for genomic predictions with the two models PSAM (population-specific effects of SNP alleles model) and ASGM (across-population SNP genotype model). pA, pB, A×B: Deli parents, La Mé parents and Deli×La Mé hybrid ortets, _(i) denotes imputed data.

II.2.1.4. 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 percentages of missing data per SNP, as shown in Table VI. 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.

II.2.1.5. Prediction models and computation of genetic values of unobserved clones

Two approaches were implemented to predict the genetic value of the validation clones: the across-population SNP genotype model (ASGM) and the population-specific effects of SNP alleles model (PSAM).

In addition, for both approaches, two models were tested: a purely additive model (ASGM_A and PSAM_A) and a model combining additive and dominance effects (ASGM_AD and PSAM_AD).

Table VI. 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

The ASGM_A approach used a model with a single random genetic effect, corresponding to the additive genetic value of the parents of the training hybrid crosses and of the validation clones. The ASGM_AD and PSAM_AD models also included a random dominance effect of crosses and ortets. The PSAM_A approach used two random effects partitioning the additive genetic values of each individual into two parts originating from Deli and La Mé alleles. All these four models were implemented separately on each trait (univariate models). For GS, the GBLUP statistical approach was used (Clark & van der Werf, 2013; Habier *et al.*, 2007), and the corresponding models were termed G_ASGM_A, G_ASGM_AD, G_PSAM_A, and G_PSAM_AD. In addition, to evaluate the usefulness of the SNP data, these four models were implemented with pedigree data instead of SNPs (control PBLUP models, termed P_ASGM_A, P_ASGM_AD, P_PSAM_A, and P_PSAM_AD).

In all cases, the models were trained with the phenotypic data of ALT hybrids and the genomic data of their parents, and the genetic values of the 42 validation clones were predicted. For all the models mentioned above, no phenotypic data of the validation clones were provided to the prediction models. This corresponds to a breeding situation where predictions are made for immature individuals (e.g., nursery plantlets belonging to crosses that were not evaluated in progeny-tests but were produced by mating the best parents selected at the end of the progeny-tests). However, ortet selection can also be made within the crosses evaluated in progeny tests.

In this case, the ortet candidates have phenotypic data records, which should be taken into consideration along with their SNP data when predicting their clonal value. This was evaluated with the G_ASGM_A model, simply including the adjusted phenotypic value of the validation ortets (see below) to the phenotypic dataset used to train 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.

II.2.1.5.1. Across-population SNP genotype models (ASGM)

The model used for the G_ASGM_AD approach was as follows:

$$y = \mathbf{X}\beta + \mathbf{Z}_1g_i + \mathbf{Z}_2g_{Deli \times LM} + \mathbf{Z}_3b + \mathbf{Z}_4p + \varepsilon$$

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), $g_i \sim N(0, \mathbf{H}_i\sigma_{a_i}^2)$ the individual additive genetic effects, $g_{Deli \times LM} \sim N(0, \mathbf{H}_{Deli \times LM}\sigma_{a_{Deli \times LM}}^2)$ the genetic dominance effects, $b \sim N(0, \mathbf{I}\sigma_b^2)$ the incomplete block effect, and $p \sim N(0, \mathbf{I}\sigma_p^2)$ the elementary plot effects. \mathbf{X} , \mathbf{Z}_1 , \mathbf{Z}_2 , \mathbf{Z}_3 and \mathbf{Z}_4 are the incidence matrices associated to β , g_i , $g_{Deli \times LM}$, b and p respectively. $\mathbf{H}_i\sigma_{a_i}^2$ and $\mathbf{H}_{Deli \times LM}\sigma_{a_{Deli \times LM}}^2$ are the variance-covariance matrices associated with g_i and $g_{Deli \times LM}$, respectively. $\sigma_{a_i}^2$ and $\sigma_{a_{Deli \times LM}}^2$ are the additive and dominance variances, respectively. $\varepsilon \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$ is the vector of residual effects and \mathbf{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 IV), and the \mathbf{H}_i 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 \mathbf{H}_i matrices subsequently in this thesis work will refer to matrices combining genealogical and genomic information. \mathbf{H}_i^{-1} is the inverse of \mathbf{H}_i , computed according to Misztal et al. (2009) 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 genealogical additive relationship matrices, respectively, of the 42 ortets and the hybrid crosses with genotyped parents, and \mathbf{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 IV). The genomic additive relationship matrix \mathbf{G} was obtained as: $\mathbf{G} = \frac{\mathbf{X}\mathbf{X}'}{2\sum_{l=1}^{n_{\text{SNP}}} p_l(1-p_l)}$, with $\mathbf{X} = \mathbf{Z} - \mathbf{P}$, \mathbf{X}' the transpose of matrix \mathbf{X} , \mathbf{Z} the SNP matrix containing the number of copies of the minor allele at an SNP (ranging from 0 to 2), \mathbf{P} a matrix given by $\mathbf{P} = 2p_l$, and p_l the frequency of the minor allele at SNP l (VanRaden, 2008). $\mathbf{H}_{\text{Deli} \times \text{LM}}$ 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. $\mathbf{H}_{\text{Deli} \times \text{LM}}^{-1}$ was computed following the same method as \mathbf{H}_i^{-1} except that the additive relationship matrices were replaced by the dominance relationship matrices. The realized dominance relationship matrix \mathbf{G}_D was computed according to Su et al. (Su et al., 2012) as: $\mathbf{G}_D = \frac{\mathbf{\Pi}\mathbf{\Pi}'}{2\sum p_l q_l (1-2p_l q_l)}$, with $\mathbf{\Pi}$ the $n \times m$ matrix (n : number of hybrid crosses and clones and m : number of SNPs) of heterozygosity coefficients 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 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, \mathbf{H}_i was replaced by the additive genealogical relationship matrix \mathbf{A}_i and, for P_ASGM_AD, $\mathbf{H}_{\text{Deli} \times \text{LM}}$ 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, $\hat{g}_i + \hat{g}_{\text{Deli} \times \text{LM}}$.

II.2.1.5.2. Population-specific effects of SNP alleles models (PSAM)

The model used for G_PSAM_AD was as follows:

$$y = \mathbf{X}\beta + \mathbf{Z}_1 g_{\text{Deli}} + \mathbf{Z}_2 g_{\text{LM}} + \mathbf{Z}_3 g_{\text{Deli} \times \text{LM}} + \mathbf{Z}_4 b + \mathbf{Z}_5 p + \varepsilon$$

with $g_{\text{Deli}} \sim N(0, \mathbf{H}_{\text{Deli}}\sigma_{g_{\text{Deli}}}^2)$ and $g_{\text{LM}} \sim N(0, \mathbf{H}_{\text{LM}}\sigma_{g_{\text{LM}}}^2)$ the additive effects inherited by the parents of the hybrid crosses and the ortets from the Deli and La Mé populations, respectively,

and $g_{Deli \times LM} \sim N(0, \mathbf{H}_{Deli \times LM} \sigma_{d_{Deli \times LM}}^2)$ the dominance effects of the crosses and clones. \mathbf{X} , $\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, \mathbf{Z}_4, \mathbf{Z}_5$ are the incidence matrices associated to $\beta, g_{Deli}, g_{LM}, g_{Deli \times LM}, b$ and p , respectively. $\mathbf{H}_{Deli} \sigma_{g_{Deli}}^2, \mathbf{H}_{LM} \sigma_{g_{LM}}^2$ and $\mathbf{H}_{Deli \times LM} \sigma_{d_{Deli \times LM}}^2$ are the variance-covariance matrices associated to 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 and La Mé populations, respectively, and $\sigma_{d_{Deli \times LM}}^2$ is the genetic dominance variance of crosses and clones. \mathbf{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. \mathbf{H}_{LM} is defined similarly for the La Mé population. To build \mathbf{H}_{Deli} , we created first the matrix of additive realized relationships of Deli parents \mathbf{G}_{Deli} (incorporating the Deli parents of the training and validation hybrid crosses and clones) as follows (Xiang *et al.*, 2016):

$$\mathbf{G}_{Deli} = \begin{bmatrix} \mathbf{G}_{Deli}^{Deli, Deli} & \mathbf{G}_{Deli}^{Deli, Deli \times LM} \\ \mathbf{G}_{Deli}^{Deli \times LM, Deli} & \mathbf{G}_{Deli}^{Deli \times LM, Deli \times LM} \end{bmatrix} \text{ with,}$$

$$\mathbf{G}_{Deli}^{Deli, Deli} = (\mathbf{Z}_{Deli} - 2\mathbf{p}_{Deli} \mathbf{1}') (\mathbf{Z}_{Deli} - 2\mathbf{p}_{Deli} \mathbf{1}'),$$

$$\mathbf{G}_{Deli}^{Deli, Deli \times LM} = (\mathbf{Z}_{Deli} - 2\mathbf{p}_{Deli} \mathbf{1}') (\mathbf{Z}_{Deli \times LM} - \mathbf{p}_{Deli} \mathbf{1}')' \text{ and}$$

$$\mathbf{G}_{Deli}^{Deli \times LM, Deli \times LM} = (\mathbf{Z}_{Deli \times LM} - \mathbf{p}_{Deli} \mathbf{1}') (\mathbf{Z}_{Deli \times LM} - \mathbf{p}_{Deli} \mathbf{1}')'.$$

\mathbf{Z}_{Deli} and $\mathbf{Z}_{Deli \times LM}$ are the matrices containing the number of copies of reference allele in the genotyped Deli parents (coded as 0, 1 or 2) and in the Deli haplotype of clones (coded as 0 or 1), respectively, \mathbf{p}_{Deli} is the vector containing the allele frequencies based on SNP genotypes of Deli parents and Deli haplotype in clones and $\mathbf{1}$ is a vector of ones. \mathbf{G}_{Deli} was then adjusted to be on the same scale and compatible with the genealogical additive relationship matrix of the clones and the genotyped Deli parents $\mathbf{A}_{Deli_{22}}$, according to Christensen *et al.* (2012) and Xiang *et al.* (2016).

\mathbf{G}_{Deli_w} and using weight 0.001, to give the \mathbf{G}_{Deli_w} matrix. Then the inverse of \mathbf{H}_{Deli} was constructed as:

$$\mathbf{H}_{Deli}^{-1} = \mathbf{A}_{Deli}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}_{Deli_w}^{-1} - \mathbf{A}_{Deli_{22}}^{-1} \end{bmatrix}, \text{ with } \mathbf{A}_{Deli}^{-1} \text{ the inverse of the genealogical relationship}$$

matrix of all the Deli parents and clones. \mathbf{H}_{LM} was created following the same procedure as \mathbf{H}_{Deli} . $\mathbf{H}_{Deli \times LM}$ 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: $\mathbf{H}_{Deli \times LM} = \mathbf{H}_{Deli} \otimes \mathbf{H}_{LM}$, with \otimes the Kronecker product.

For P_PSAM_A and P_PSAM_AD, \mathbf{H}_{Deli} and \mathbf{H}_{LM} were replaced by the additive genealogical relationship matrices \mathbf{A}_{Deli} and \mathbf{A}_{LM} and, for P_PSAM_AD, $\mathbf{H}_{Deli \times LM}$ was replaced by the genealogical dominance relationship matrix.

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

II.2.1.6. Prediction accuracies

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

Pairwise comparisons of prediction accuracies among models were made for each trait using the Hotelling–Williams *t*-test (Steiger, 1980). This test compares two non-independent correlations, i.e., having one variable in common, which in our case is the reference value of the 42 clones. This test was applied using the R package *psych* (Revelle, 2018). The Hotelling–Williams *t*-test is given as:

$$t = (r_{12} - r_{13}) \sqrt{\frac{(n-1)(1+r_{23})}{2\left(\frac{n-1}{n-3}\right)|R| + \left(\frac{r_{12}+r_{13}}{4}\right)^2(1-r_{23})^3}}$$

with $|R| = (1 - r_{12}^2 - r_{13}^2 - r_{23}^2) + (2r_{12}r_{13}r_{23})$, *t* is the *t* statistic on $(n - 1)$ degree of freedom, *n* (42 clones) the sample size, r_{12} and r_{13} are the coefficients of correlation whose differences are tested, r_{23} is the coefficient of correlation between the two predictors, $|R|$ is the determinant of the correlation matrix.

The p-values which show the significance are deducted from the obtained *t*-values.

II.2.1.7. Determination of the reference clonal values predicted by the model

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

II.2.1.8. Accuracy of phenotypic selection before clonal trials

To evaluate the possibility of using GS instead of the current phenotypic selection (PS) to select the hybrid individuals to test in the clonal trials, the PS accuracy was computed for each trait. It 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.

II.2.2. Effect of the genotyping strategy to optimize prediction accuracy

II.2.2.1. Phenotyping

Phenotypic data were collected from the hybrid individuals on nine traits, comprising BN, FFB, ABW, AFW, NF, FB, PF, OP and OER. These components were measured in palms aged from three to seven years old for bunch production and from five to nine years old for bunch quality.

II.2.2.2. Generation of SNP molecular data

A genotyping-by-sequencing (GBS) (Elshire *et al.*, 2011) approach was used to generate the SNP data of the parents of groups A and B of the training and validation hybrid crosses and of a set of 399 hybrid individuals sampled among the training crosses (Table VII and Table VIII).

Table VII. Composition of training and validation sets.

	Training population		Validation population	
	bunch production	bunch quality	bunch production	bunch quality
Number of crosses	352	341	213	213
Number of individuals (genotyped)	22,656	14,985	13,399	10,339
Average number of individuals per cross (min–max)	64 (17-503)	44 (21-292)	63 (25-680)	48 (19-493)
Number of parents in group A	123	121	71	71
Number of parents in group B	121	118	49	49
Age of trees at time of data collection (years)	3-7	5-9	3-7	5-9

The genotyped hybrid individuals belonged to 97 crosses involving respectively 59 parents of group A and 60 parents of group B (Table VIII). DNA extraction and genotyping protocol were performed as described above (II.2.1.2), yielding to a marker density of 21,458 SNPs.

II.2.2.3. Imputation of missing SNP genotypes and phasing

The initial raw SNP dataset in the form of variant call format (VCF) included parents of groups A and B of the training and validation sets and the 399 hybrid individuals of the training set having their two parents A and B genotyped. This initial dataset has been divided using VCFtools (Danecek et al., 2011) into three distinct sub-datasets i.e. one containing only the parents of group A, another with the parents of group B and the third with A×B hybrid individuals. SNP datasets of parents were imputed separately using their respective pedigrees, then merged with the unimputed SNP dataset of hybrid individuals and the whole dataset was imputed and phased using the global pedigree indicating parents A and B of hybrid individuals (see II.2.1.3).

II.2.2.4. Models for prediction of hybrid performances

Two different modeling approaches have been applied to predict the genetic value of oil palm yield components: ASGM and PSAM. Only purely additive models were considered here given that previous studies showed that modeling dominance effects did not improve predictive abilities (Cros et al., 2017; Nyouma et al., 2020). Additive genetic values of the validation crosses were predicted using SNP molecular data for both approaches aforementioned thus becoming G_ASGM_Par and G_PSAM_Par.

In addition in each model, the effect of adding molecular hybrid individual information has been assessed. These models were termed ASGM_Par+Hyb and G_PSAM_Par+Hyb.

To confirm the usefulness SNPs, the control, pedigree-based approach, were assessed with similar models termed as P_PSAM_Par and P_ASGM_Par, and when hybrid individuals were present P_PSAM_Par+Hyb and P_ASGM_Par+Hyb.

Table VIII. Characteristics of genotyped hybrid individuals of the training set.

	bunch production	bunch quality
Number of hybrid individuals	397	399
Number of crosses	97	97
Average number of individuals per cross (min–max)	4 (1-10)	4 (1-10)
Number of group A parents	59	59
Number of group B parents	60	60

II.2.2.4.1. Across-population SNP genotype models (ASGM)

The approach ASGM considers that the allele effect of a given SNP is the same for group A or group B. For this reason, ASGM contains a single genetic effect, g_g , corresponding to the additive genetic value of the parents of the training and validation crosses (ASGM_Par), or of the parents of the training and validation crosses and the training hybrid individuals (ASGM_Par+Hyb). The G_ASGM models were as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_g\mathbf{g}_g + \mathbf{Z}_b\mathbf{b} + \mathbf{Z}_p\mathbf{p} + \boldsymbol{\varepsilon}$$

where \mathbf{y} is the vector of hybrid phenotypes (BN, ABW, FFB, AFW, FB, PF, OP, OER or NF) of the training set, β is the vector of fixed effects (overall mean of phenotypes, trial effects, block effects and, for bunch production traits, age), $\mathbf{g}_g \sim N(0, \mathbf{G}_g\sigma_{a_g}^2)$ is the vector of additive genetic effects, $\mathbf{b} \sim N(0, \mathbf{I}\sigma_b^2)$ is the vector of incomplete block effects, and $\mathbf{p} \sim N(0, \mathbf{I}\sigma_p^2)$ is the vector of the elementary plot effects. \mathbf{X} , \mathbf{Z}_g , \mathbf{Z}_b and \mathbf{Z}_p are the incidence matrices associated to vectors β , \mathbf{g}_g , \mathbf{b} and \mathbf{p} respectively. $\mathbf{G}_g\sigma_{a_g}^2$ is the variance-covariance matrix associated with \mathbf{g}_g . \mathbf{G}_g is the genomic additive relationship matrix obtained as $\mathbf{G}_g = \frac{\mathbf{X}\mathbf{X}'}{2\sum_{l=1}^n p_l q_l}$, with $\mathbf{X} = \mathbf{Z} - \mathbf{P}$; \mathbf{X}' is the transpose of matrix \mathbf{X} ; \mathbf{Z} is the matrix of SNP containing for each SNP the number of copies of the reference allele coded into 0, 1 and 2; $\mathbf{P} = 2\mathbf{p}_l$ is a matrix with p_l the frequency of the minor allele at SNP l and $q_l = (1 - p_l)$ (VanRaden, 2008). $\sigma_{a_g}^2$ is the additive variance. $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$ is the vector of residual effects and \mathbf{I} the identity matrix.

Among the phenotyped hybrid individuals, 399 were genotyped. For the ungenotyped hybrid individuals, the expected genotype was computed based on the genotypes of their two parents in groups A and B, i.e., the number of copies of reference alleles in ungenotyped hybrid individuals of a given cross was considered equal to the mean number of copies of the reference alleles of their parents.

For G_ASGM_Par, the \mathbf{G}_g matrix only contained the hybrid parents. For G_ASGM_Par+Hyb, the \mathbf{G}_g matrix contained the hybrid parents and the hybrid individuals.

Based on the results obtained in preliminary analyses, the \mathbf{G}_g matrices were adjusted according to the method described in Christensen *et al.* (2014) and Xiang *et al.* (2016), with the α and β adjustment parameters estimated from the genomic and genealogical data and ω taken as 0.001.

For P_ASGM_Par+Hyb and P_ASGM_Par, \mathbf{G}_g matrices were replaced by genealogical relationship matrices.

The estimated additive genetic value of the validation crosses was found in the \hat{g}_g vector.

II.2.2.4.2. Population-specific effects of SNP alleles models (PSAM)

The PSAM model distinguishes the parental origin of group A and B alleles. Consequently, PSAM comprises two distinct genetic effects, g_A and g_B , corresponding to the additive genetic value of parents A and B, respectively, and, for the hybrids, to the additive value resulting from the alleles inherited from parents A and B, respectively. The PSAM models were as follows:

$$y = X\beta + Z_A g_A + Z_B g_B + Z_b b + Z_p p + \varepsilon$$

where $g_A \sim N(0, G_A \sigma_{aA}^2)$ and $g_B \sim N(0, G_B \sigma_{aB}^2)$, and Z_A and Z_B are the incidence matrices associated to vectors g_A and g_B , respectively. $G_A \sigma_{aA}^2$ and $G_B \sigma_{aB}^2$ are the variance-covariance matrices associated to g_A and g_B . G_A and G_B are the matrices of additive realized relationships of groups A and B (Fig. 18. and Fig. 19). For G_PSAM_Par, they G_A and G_B included the parents of the training and validation hybrid crosses and, for G_PSAM_Par+Hyb, they also contained the training hybrid individuals. G_PSAM_Par, G_A and G_B were calculated similarly as G_g in ASGM models (VanRaden, 2008). However, for G_PSAM_Par+Hyb i.e. including hybrid individuals G_A and G_B were constructed according Christensen *et al.* (2014) and Xiang *et al.* (2016) as follow:

$$G_A = \begin{bmatrix} G_A^{A,A} & G_A^{A,AB} \\ G_A^{AB,A} & G_A^{AB,AB} \end{bmatrix} \text{ with:}$$

$$G_A^{A,A} = (Z_A - 2P_A) (Z_A - 2P_A)',$$

$$G_A^{A,AB} = (Z_A - 2P_A) (Z_{AB} - P_A)' \text{ and}$$

$$G_A^{AB,AB} = (Z_{AB} - P_A) (Z_{AB} - P_A)'.$$

Where Z_A and Z_{AB} are the matrices containing the reference population-specific alleles of the parents A (coded as 0, 1 or 2) and the number of copies of the reference allele phase of the parents A of A×B hybrid individuals (coded as 0 or 1), p_A is the vector containing the specific allele frequencies based on SNP genotypes for parents A and specific SNP allele of parents A for A×B hybrid individuals and $\mathbf{1}$ is a vector of ones.

G_A is then adjusted to be on the same scale and compatible with the genealogical relationship matrix A_A .

$\mathbf{G}_{A_a} = \mathbf{G}_A \beta + \alpha$. The parameters α and β are unknown and can be estimated by solving the following system of equations:

$$\begin{cases} \overline{d\mathbf{G}_A} \beta + \alpha = \overline{d\mathbf{A}_A} \\ \overline{\mathbf{G}_A} \beta + \alpha = \overline{\mathbf{A}_A} \end{cases}$$

$\overline{d\mathbf{G}_A}$ and $\overline{d\mathbf{A}_A}$ are the averages of diagonals of matrices \mathbf{G}_A and \mathbf{A}_A , $\overline{\mathbf{G}_A}$ and $\overline{\mathbf{A}_A}$ are the averages of the matrices \mathbf{G}_A and \mathbf{A}_A .

The solutions of this equation system are:

$$\beta = (\overline{\mathbf{A}_A} - \overline{d\mathbf{A}_A}) / (\overline{\mathbf{G}_A} - \overline{d\mathbf{G}_A})$$

$$\alpha = (\overline{d\mathbf{A}_A} - \overline{d\mathbf{G}_A}) \times (\overline{\mathbf{A}_A} - \overline{d\mathbf{A}_A}) / (\overline{\mathbf{G}_A} - \overline{d\mathbf{G}_A})$$

\mathbf{G}_{A_a} is then adjusted to integrate the part of genetic variance not captured by the SNPs.

$\mathbf{G}_{A_w} = (1 - w)\mathbf{G}_{A_a} + w\mathbf{A}_A$. With the parameter w a relative constant giving the proportion of genetic variance that is not captured by the SNPs. Many values have been assessed and 0.001 have been chosen because it was maximizing prediction accuracies and minimizing biases for hybrid individuals.

For group A, matrices \mathbf{G}_B , \mathbf{G}_{B_a} and \mathbf{G}_{B_w} were computed similarly with \mathbf{G}_A , \mathbf{G}_{A_a} and \mathbf{G}_{A_w} .

For P_ASGM_Par+Hyb and P_ASGM_Par, \mathbf{G}_{A_w} and \mathbf{G}_{B_w} matrices were replaced by the genealogical relationship matrices \mathbf{A}_A and \mathbf{A}_B .

The estimated additive genetic value of the validation crosses was the sum of the genetic additive value inherited from the parent of groups A and B, i.e., $\hat{g}_A + \hat{g}_B$.

II.2.2.5. Reference genetic values of hybrid crosses

For each trait, the true estimated genetic value of the validation hybrid crosses, termed reference genetic values, was computed from the phenotypic data of their hybrid individuals using a linear mixed model in which the overall mean of hybrid crosses, cross effects, trial effects, block effects, and for bunch production, age, have been used as fixed effects and hybrid individuals, elementary plots and incomplete blocs as random effects.

II.2.2.6. Prediction accuracies and model comparison of models

The prediction accuracies have been computed for each trait and each model as the correlation between the reference genetic values and the genetic values of the training crosses.

To compare the prediction accuracy of the models, the 213 A×B validation crosses were divided into eight replicates, i.e., five replicates made up of 27 crosses and three replicates made

up of 26 crosses. Each of the parents involved in these validation crosses was genotyped and none were parents involved in training crosses. The prediction accuracies were computed for each replicate, allowing the comparison of models using an analysis of variance (ANOVA) or the Wald-type permutation test of the R package GFD (Friedrich et al., 2017) when the normality of residuals and the homoscedasticity assumptions were not met. These statistical tests were implemented using the *agricolae* R package (Mendiburu, 2016).

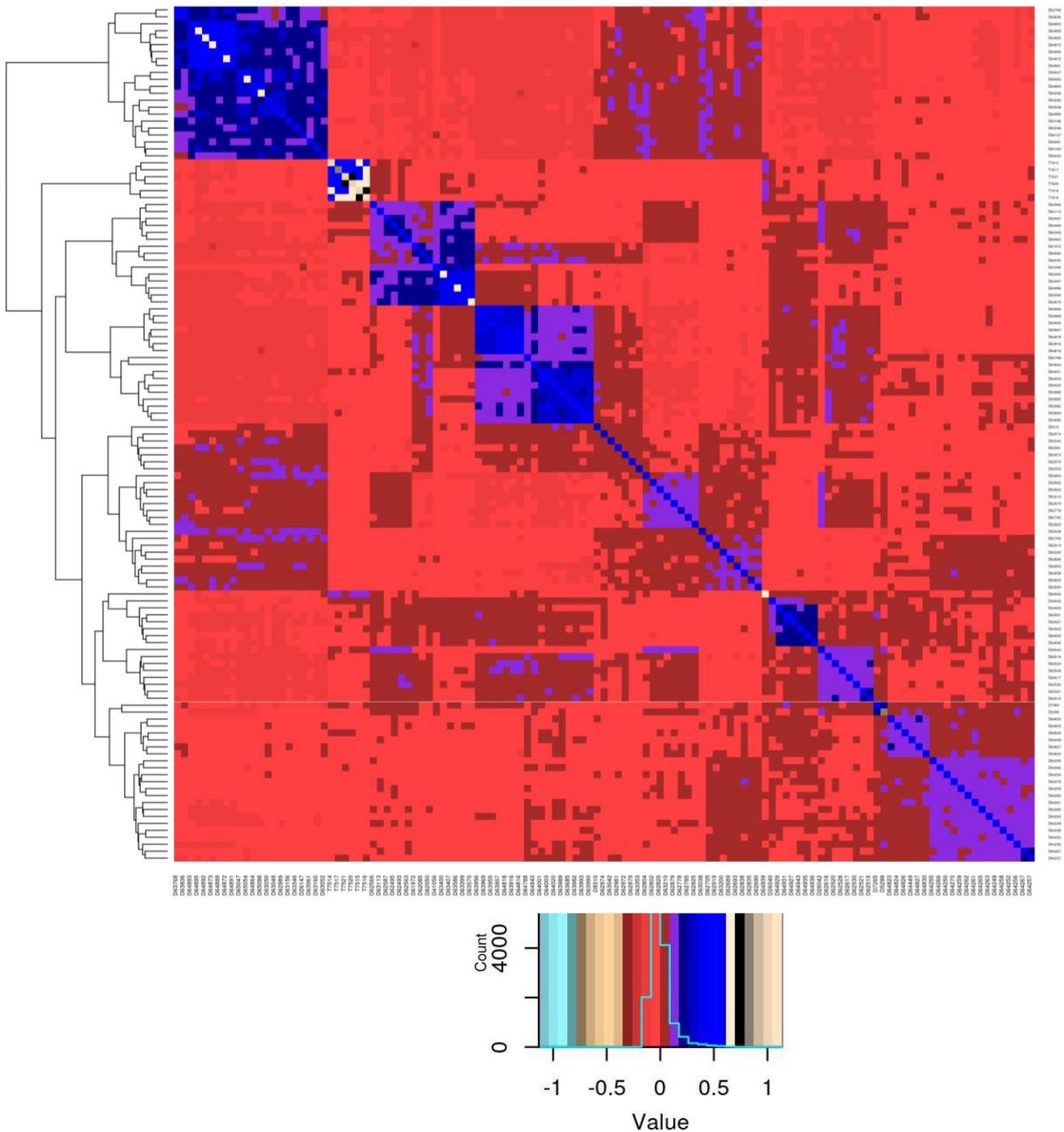


Fig. 18. Heat map of additive realized relationships matrices of the 123 parents A of the training set.

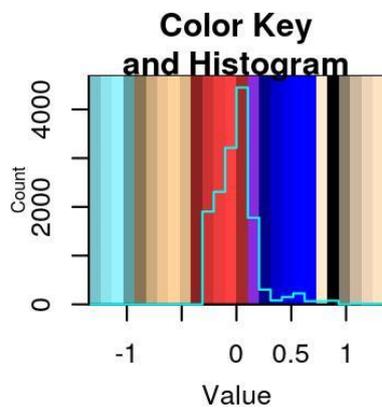
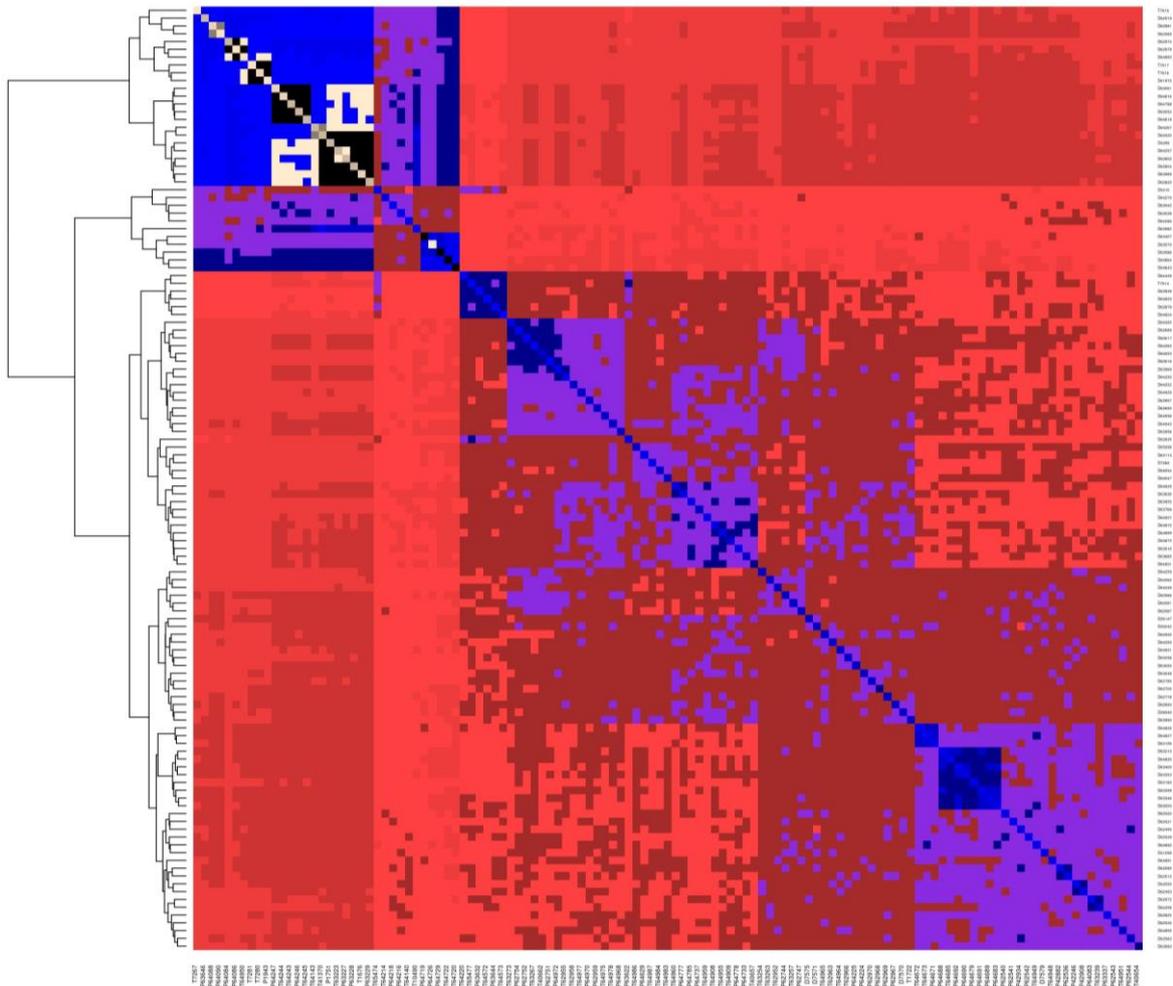


Fig. 19. Heat map of additive realized relationships matrices of the 121 parents B of the training set.

CHAPTER III. RESULTS AND DISCUSSION

III.1. Results

III.1.1. Efficiency of genomic selection for clonal selection

III.1.1.1. 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. 20).

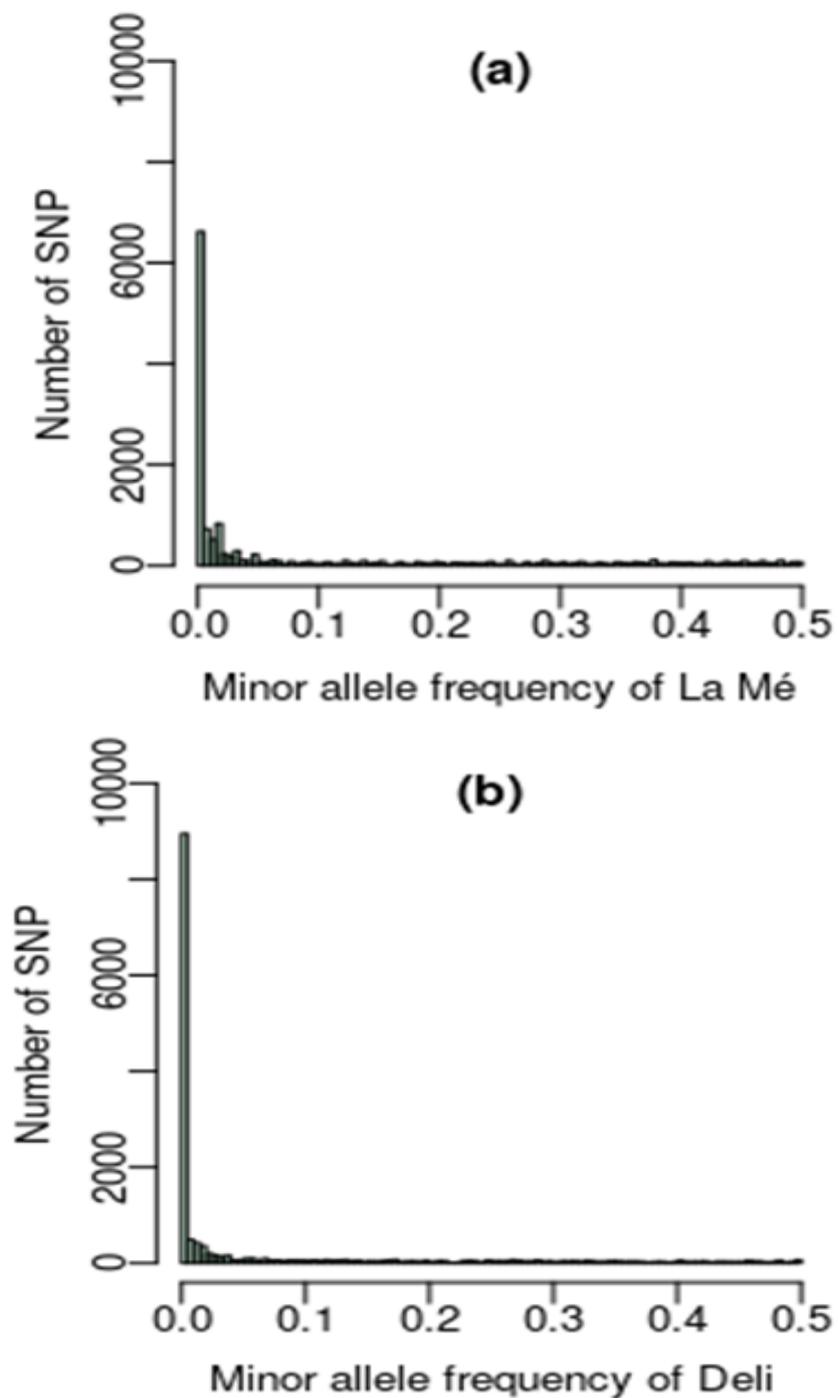


Fig. 20. Distribution of minor allele frequency (MAF). a: La Mé population; b: Deli population.

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. 20a) and 0.07 for Deli (Fig. 20b). 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. 21a) 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 $MAF \geq 0.05$ in the other population. This is the result of the high genetic difference between Deli and La Mé populations, for which the F_{st} fixation index reaches 0.55 (Cros *et al.*, 2018). 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 $MAF \geq 0.05$ in La Mé) than in La Mé (11.9% of all SNPs had $MAF < 0.05$ in La Mé and $MAF \geq 0.05$ 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. 21b) 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 $MAF \geq 0.05$ 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.

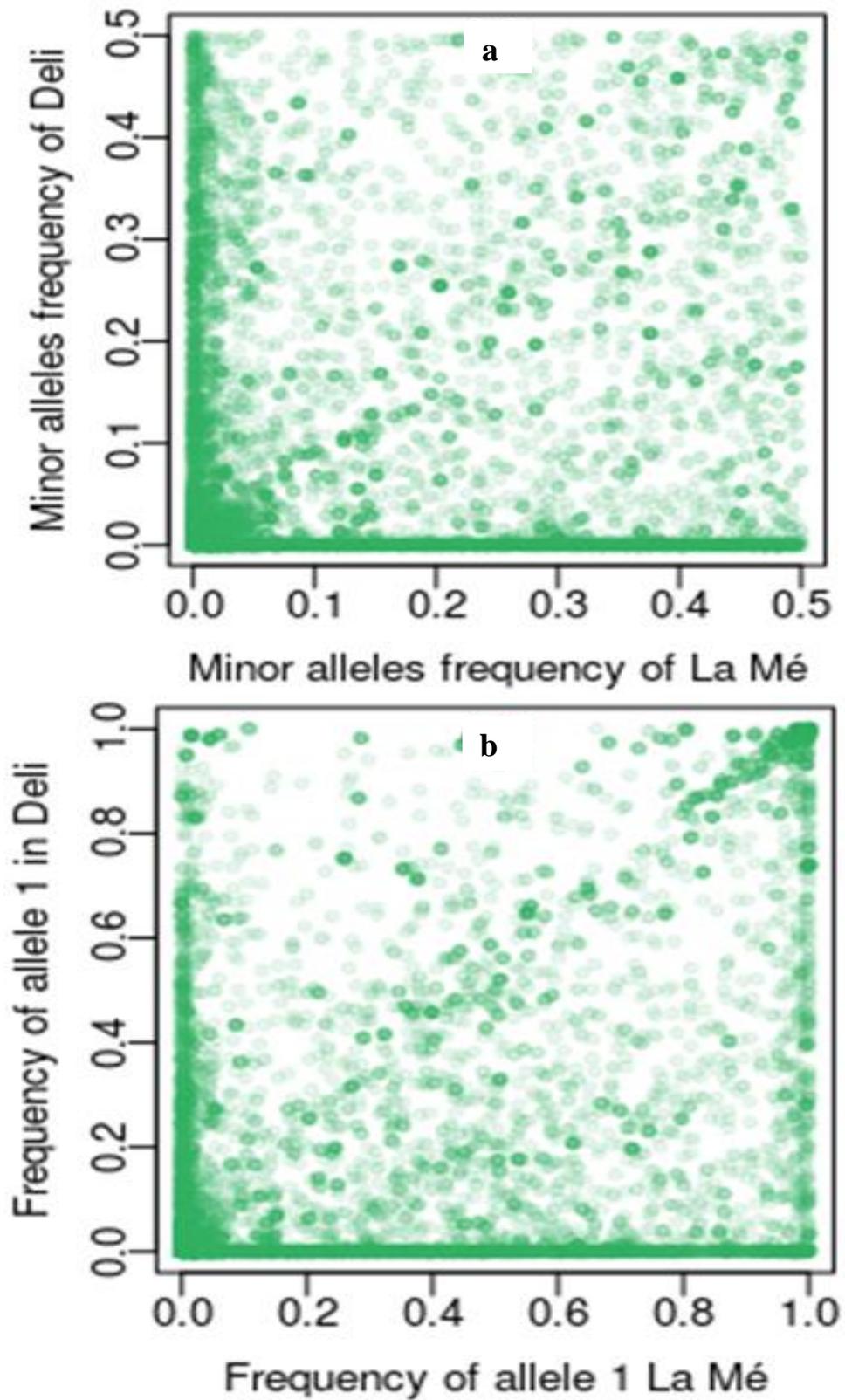


Fig. 21. Correlation of minor allele frequency (MAF) (a) and frequency of alternate alleles between La Mé and Deli (b) populations. In (a) and (b) panels, each dot represents an SNP.

III.1.1.2. Effect of GS prediction model and SNP dataset on prediction accuracy

Prediction accuracies of GS methods ranged from -0.04 to 0.70 depending on the prediction model, trait and SNP dataset.

Genomic prediction accuracies of additive + dominance models ranged from -0.04 to 0.66 depending on trait, prediction model and SNP dataset (Fig. 22). Prediction accuracies of GS of additive + dominance models for bunch production traits ranged from 0.1 to 0.62 depending on model and SNP dataset. Firstly, for G_ASGM_AD models, prediction accuracies increased with SNP dataset up to the SNP dataset $p_{max}=10\%-n_{SNP}=6,898$ where it plateaued for both BN and ABW (Fig. 22a, b) and started to slightly decrease for FFB. Prediction accuracies of G_ASGM_AD models ranged from 0.25 ($p_{max}=0\%-n_{SNP}=2,447$) to 0.35 ($p_{max}=75\%-n_{SNP}=15,054$) for BN (Fig. 22a), 0.39 ($p_{max}=0\%-n_{SNP}=2,447$) to 0.53 ($p_{max}=75\%-n_{SNP}=15,054$) for ABW (Fig. 22b) and 0.2 ($p_{max}=0\%-n_{SNP}=2,447$) to 0.39 ($p_{max}=10\%-n_{SNP}=6,898$) for FFB (Fig. 22c). Secondly, For G_PSAM_AD models, prediction accuracies increased in general with SNP dataset as in G_ASGM_AD models although varied considerably with the dataset for FFB (Fig. 22a, b, c). Prediction accuracies extended from 0.1 at the SNP dataset $p_{max}=5\%-n_{SNP}=5,620$ to 0.28 at SNP datasets $p_{max}=25\%-n_{SNP}=9,059$ and $p_{max}=75\%-n_{SNP}=15,054$ for BN (Fig. 22a), 0.52 at the SNP dataset $p_{max}=0\%-n_{SNP}=2,447$ to 0.62 at SNP datasets $p_{max}=5\%-n_{SNP}=5,620$ and $p_{max}=10\%-n_{SNP}=6,898$ for ABW (Fig. 22b) and 0.22 at the SNP dataset $p_{max}=0\%-n_{SNP}=2,447$ to 0.55 at the SNP dataset $p_{max}=45\%-n_{SNP}=11,425$ for FFB (Fig. 22c).

Prediction accuracies of bunch quality traits extended from -0.04 to 0.66 depending on model, trait and SNP dataset. Firstly, for G_ASGM_AD models, prediction accuracies varied widely with the SNP dataset in an inconsistent way. Prediction accuracies extended from 0.41 ($p_{max}=75\%-n_{SNP}=15,054$) to 0.55 ($p_{max}=0\%-n_{SNP}=2,447$) for AFW (Fig. 23a), 0.4 ($p_{max}=0\%-n_{SNP}=2,447$, $p_{max}=45\%-n_{SNP}=11,425$ and $p_{max}=75\%-n_{SNP}=15,054$) to 0.47 ($p_{max}=25\%-n_{SNP}=9,059$) for FB (Fig. 23b), -0.04 ($p_{max}=0\%-n_{SNP}=2,447$) to 0.11 ($p_{max}=45\%-n_{SNP}=11,425$) for PF (Fig. 23c), 0.38 ($p_{max}=5\%-n_{SNP}=5,620$) to 0.50 ($p_{max}=0\%-n_{SNP}=2,447$) for OP (Fig. 23d) and 0.45 ($p_{max}=45\%-n_{SNP}=11,425$) to 0.6 ($p_{max}=75\%-n_{SNP}=15,054$) for NF (Fig. 23e). Secondly, for G_PSAM_AD models, prediction accuracies considerably varied depending on the SNP dataset in an inconsistent way as in G_ASGM_AD. Prediction accuracies ranged from 0.46 ($p_{max}=0\%-n_{SNP}=2,447$ and $p_{max}=25\%-n_{SNP}=9,059$) to 0.50 ($p_{max}=5\%-n_{SNP}=5,620$ and $p_{max}=10\%-n_{SNP}=6,898$) for AFW (Fig. 23a), 0.53 ($p_{max}=45\%-n_{SNP}=11,425$) to 0.66 ($p_{max}=10\%-n_{SNP}=6,898$) for FB (Fig. 23b), 0.12 ($p_{max}=5\%-n_{SNP}=5,620$ and $p_{max}=45\%-n_{SNP}=11,425$) to 0.26 ($p_{max}=10\%-n_{SNP}=6,898$) for PF (Fig. 23c), 0.37 ($p_{max}=45\%-n_{SNP}=11,425$ and $p_{max}=75\%-n_{SNP}=15,054$) to

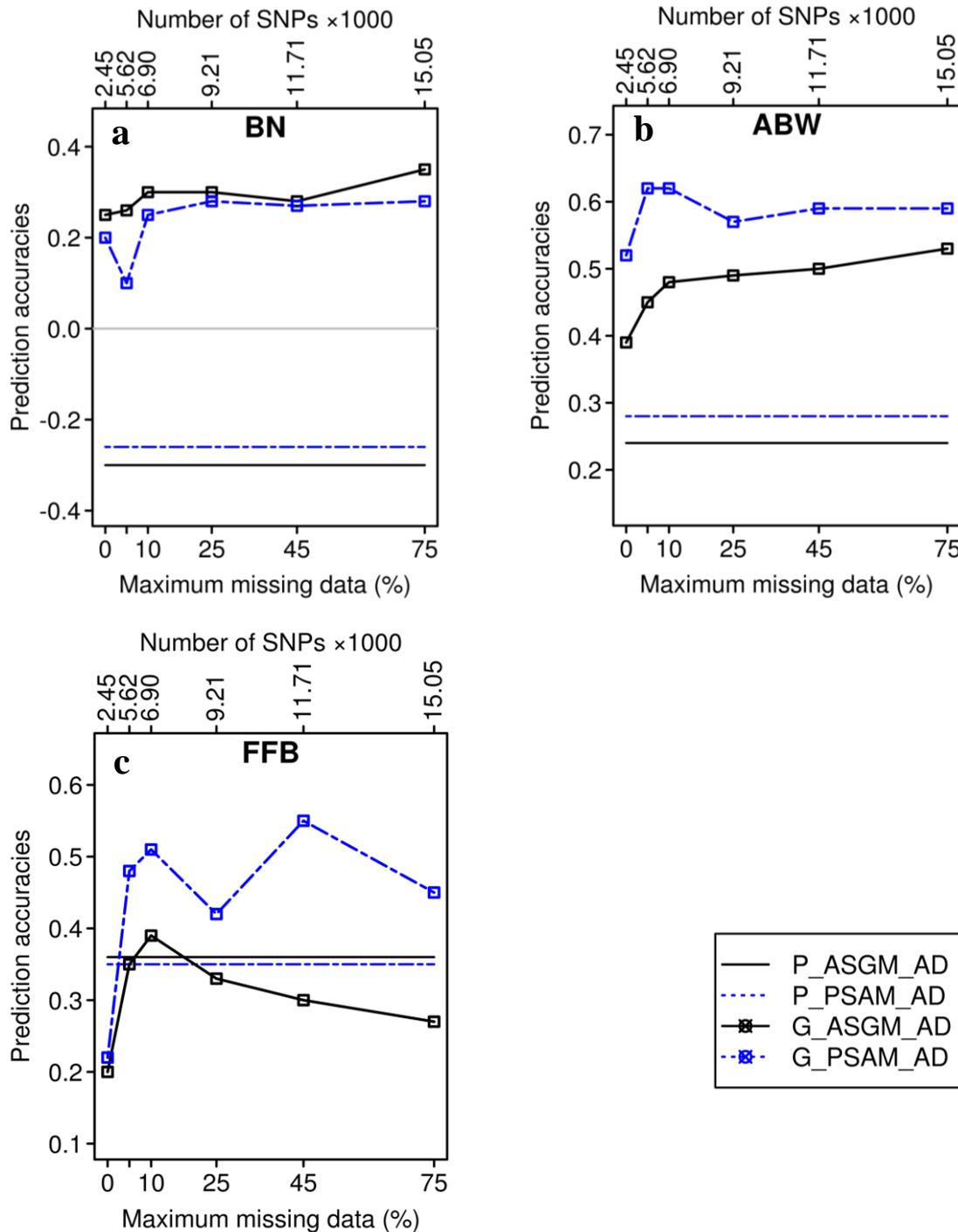


Fig. 22. Prediction accuracies of bunch production traits according to SNP datasets and prediction models.

a: bunch number (BN); b: average bunch weight (ABW); c: total bunch production (FFB). Pedigree-based prediction models: across-population SNP genotype models (P_ASGM_AD), population-specific effects of SNP alleles models (P_PSAM_AD); additive + dominance genomic prediction models: across-population SNP genotype models (G_ASGM_AD), population-specific effects of SNP alleles models (G_PSAM_AD).

0.47 ($p_{max}=10\%-n_{SNP}=6,898$) for OP (Fig. 23d) and 0.52 ($p_{max}=25\%-n_{SNP}=9,059$) to 0.61 ($p_{max}=10\%-n_{SNP}=6,898$) for NF (Fig. 23e). These analyses depicted inconsistent differences or similar accuracies between additive models and additive + dominance models, depending on SNP dataset and trait. Henceforward, we will only refer to additive models.

Prediction accuracies of GS methods ranged from 0.08 to 0.70 depending on the prediction model, trait and SNP dataset (Fig. 24 and Fig. 25) for purely additive models (G_ASGM_A and G_PSAM_A).

For bunch production components, GS prediction accuracy ranged from, 0.20 to 0.63 depending on trait and SNP dataset (Fig. 24). Prediction accuracies of GS for BN ranged from 0.20 to 0.40 depending on the model and SNP dataset (Fig. 24a). GS prediction accuracies for both modeling approaches, G_PSAM_A and G_ASGM_A increased in general from around 0.2 to 0.37 where they seemed to plateau. Genomic prediction accuracies of G_ASGM_A and G_PSAM_A models were not significantly different in all the SNP datasets considered. The highest genomic prediction accuracy for BN was observed at the SNP dataset $p_{max}=75\%-n_{SNP}=15,054$ for both GS modeling approaches G_ASGM_A and G_PSAM_A (0.35 and 0.37, respectively). Regarding the pedigree-based models, P_PSAM_A was significantly different than P_ASGM_A, with respective prediction accuracies of - 0.05 and - 0.3 (Fig. 24).

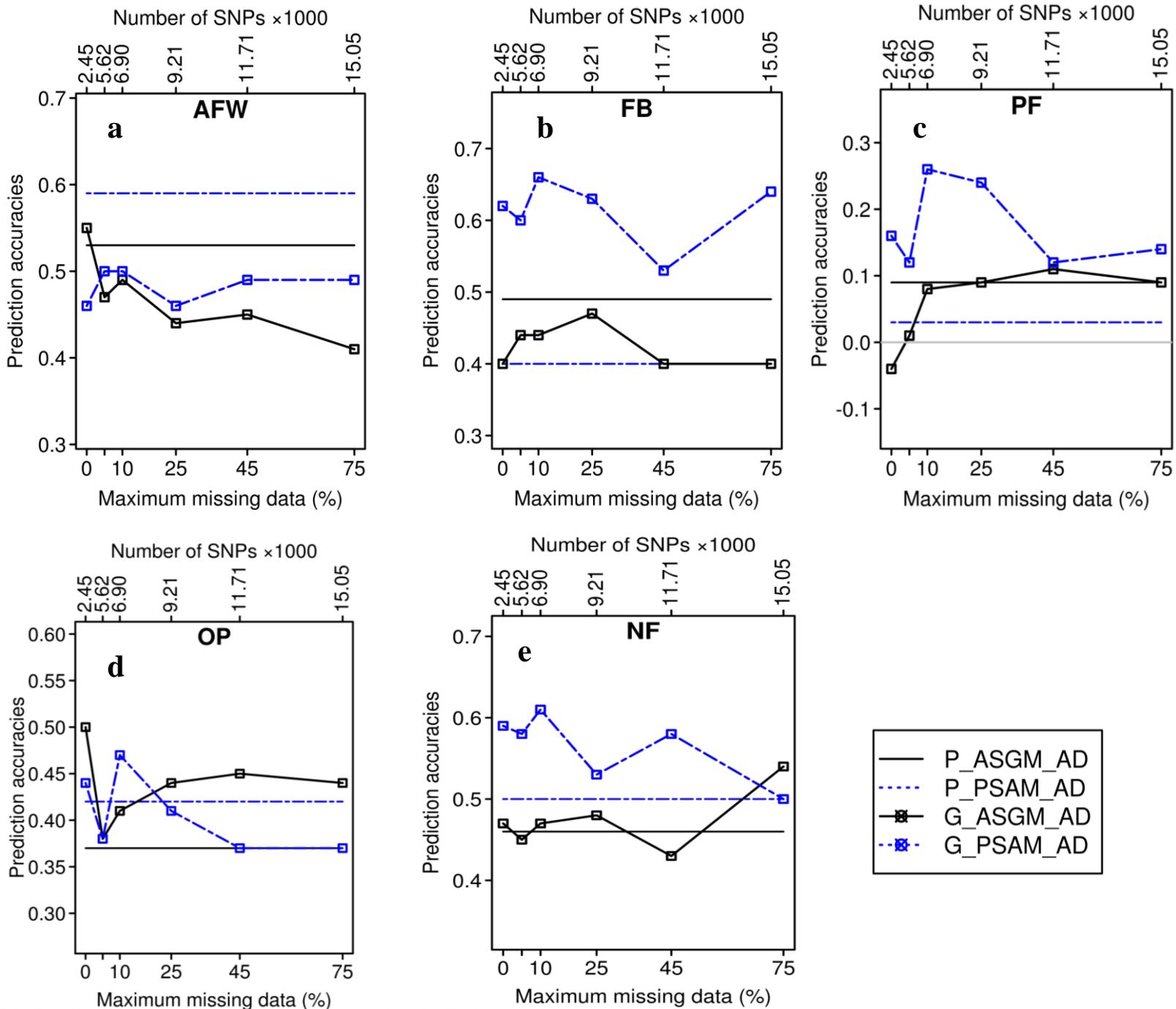


Fig. 23. Prediction accuracies according to traits, SNP datasets and prediction models.

a: average fruit weight (AFW); b: fruit to bunch (FB) ratio; c: pulp to fruit (PF) ratio; d: oil to pulp (OP) ratio; e: number of fruits (NF) per bunch. Pedigree-based prediction models: across-population SNP genotype models (P_ASGM_AD), population-specific effects of SNP alleles models (P_PSAM_AD); additive + dominance genomic prediction models: across-population SNP genotype models (G_ASGM_AD), population-specific effects of SNP alleles models (G_PSAM_AD).

Prediction accuracies of GS for ABW ranged from 0.43 to 0.63 according to the model and the SNP dataset (Fig. 24b). For the two genomic prediction approaches (G_PSAM_A and G_ASGM_A), the accuracy increased at the SNP dataset $p_{max}=0\%-n_{SNP}=2,447$ with 0.43 to 0.55 and then started to decrease until the SNP dataset $p_{max}=25\%-n_{SNP}=9,205$ with prediction accuracy of 0.56. Subsequently, genomic prediction accuracies increased up to 0.59 at the SNP dataset $p_{max}=75\%-n_{SNP}=15,054$. For G_ASGM_A, prediction accuracy increased from 0.43 to around 0.58 where it plateaued at the SNP dataset $p_{max}=10\%-n_{SNP}=6,898$ and then slightly increased. No significant difference was observed between G_ASGM_A and G_PSAM_A for all the SNP datasets. Regarding the pedigree-based model, P_PSAM was significantly higher than P_ASGM_A, with respective prediction accuracies of 0.29 and 0.24 (Fig. 24b).

Genomic prediction accuracies of FFB ranged from 0.24 to 0.55, depending on the SNP dataset and the modeling approach (Fig. 24c). For, both prediction modeling, G_ASGM_A and G_PSAM_A, prediction accuracy highly increased from the SNP dataset $p_{max}=0\%-n_{SNP}=2,447$ to $p_{max}=10\%-n_{SNP}=6,898$, then slightly decrease for G_ASGM_A. Concerning the pedigree-based model, P_ASGM_A with prediction accuracy of 0.36 was higher than P_PSAM_A with 0.33 although not significant.

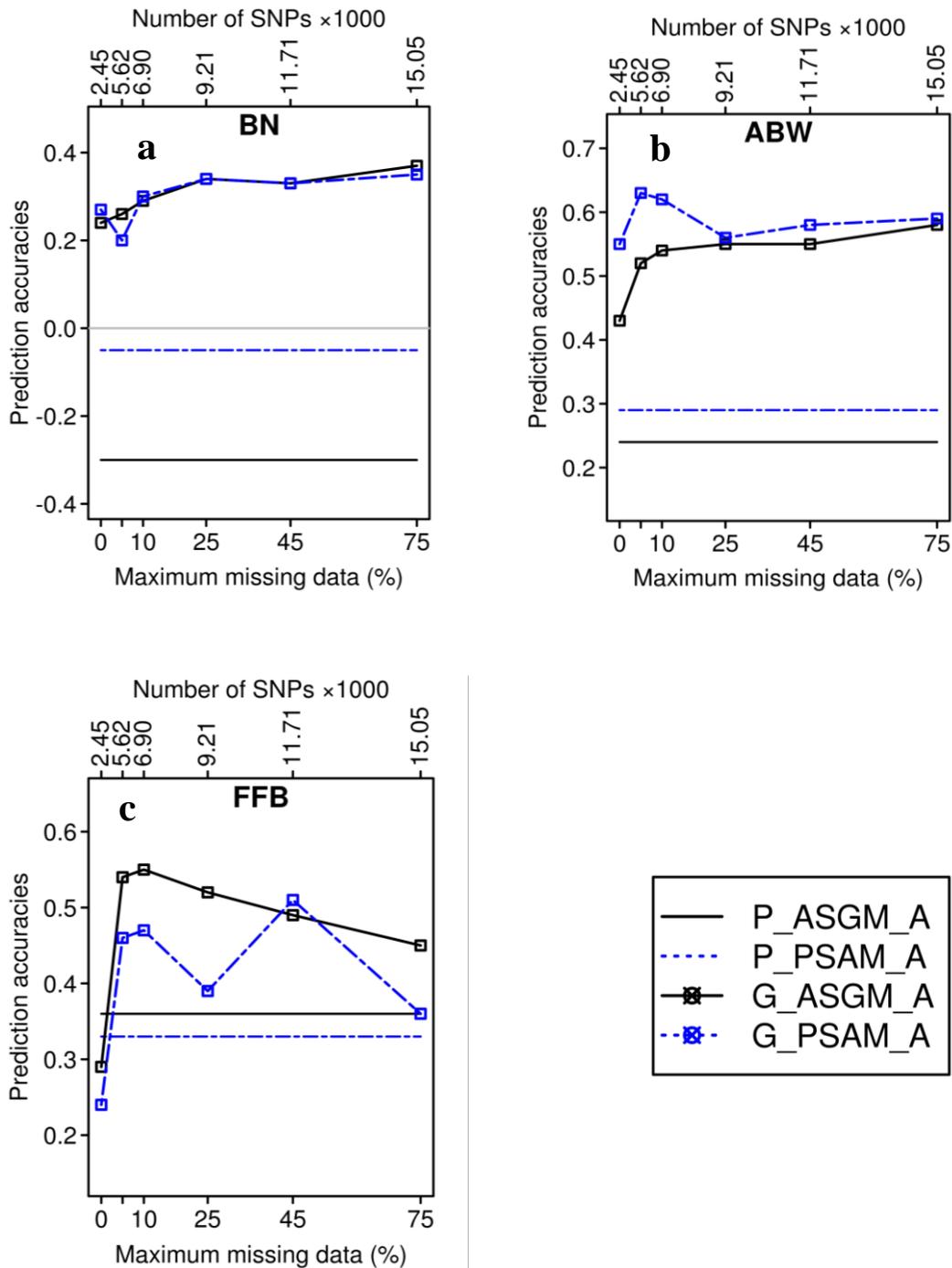


Fig. 24. Prediction accuracies of bunch production traits according to SNP datasets and prediction models.

a: bunch number (BN); b: average bunch weight (ABW); c: total bunch production (FFB). Pedigree-based prediction models: across-population SNP genotype models (P_ASGM_A), population-specific effects of SNP alleles models (P_PSAM_A); additive genomic prediction models: across-population SNP genotype models (G_ASGM_A), population-specific effects of SNP alleles models (G_PSAM_A).

Prediction accuracies of GS methods ranged from 0.08 to 0.7 for bunch quality traits, depending on the SNP dataset and prediction modeling approach (Fig. 25).

Genomic prediction accuracies for AFW ranged from 0.42 to 0.57 depending on the prediction model and the SNP dataset (Fig. 25a). For G_PSAM_A, prediction accuracies increased from 0.43 at the SNP dataset $p_{max}=0\%-n_{SNP}=2,447$ to 0.51 at the SNP dataset $p_{max}=10\%-n_{SNP}=6,898$, then decreased to 0.45 at the SNP dataset $p_{max}=25\%-n_{SNP}=9,205$ where it increased again and plateaued afterwards. For the same trait, prediction accuracies of G_ASGM_A decreased from 0.57 at the SNP dataset $p_{max}=0\%-n_{SNP}=2,447$ to 0.42 at the SNP dataset $p_{max}=45\%-n_{SNP}=11,707$ and stabilised afterwards. Regarding the pedigree-based models, prediction accuracy of P_PSAM_A i.e., 0.58 was higher than prediction accuracy of P_ASGM_A with 0.53 (Fig. 25a).

Prediction accuracies of GS for FB ranged from 0.49 to 0.7 depending on the SNP dataset (Fig. 25b). Prediction accuracies of FB for G_ASGM_A increased from 0.61 at the SNP dataset $p_{max}=0\%-n_{SNP}=2,447$ to 0.7 at the SNP dataset $p_{max}=25\%-n_{SNP}=9,205$, then slightly decreased thereafter (Fig. 25b). Concerning G_PSAM_A, prediction accuracies overall increased for the three first SNP datasets, then started to decrease and increase again afterwards. A significant difference was observed between the prediction accuracy of P_ASGM_A (0.49) and P_PSAM_A (0.35) (Fig. 25b).

For PF, GS prediction accuracies ranged from 0.08 to 0.23 depending on the SNP dataset (Fig. 25c). For G_PSAM_A, the accuracy increased from 0.08 ($p_{max}=0\%-n_{SNP}=2,447$) to 0.23 ($p_{max}=10\%-n_{SNP}=6,898$) then decrease and stabilized at 0.1 ($p_{max}=45\%-n_{SNP}=11,707$). Prediction accuracies of G_ASGM_A increased from 0.9 at the SNP dataset $p_{max}=0\%-n_{SNP}=2,447$ to 0.16 at the SNP dataset $p_{max}=10\%-n_{SNP}=6,898$, where it plateaued. The pedigree-based models showed small prediction accuracies for P_PSAM_A and P_ASGM_A models, i.e., 0.03 and 0.09, respectively.

Prediction accuracies of GS for OP ranged from 0.33 to 0.55 according to the SNP dataset (Fig. 25c). For G_PSAM_A, prediction accuracy firstly decreased from 0.41 ($p_{max}=0\%-n_{SNP}=2,447$) to 0.36 ($p_{max}=5\%-n_{SNP}=5,620$) then increased up to a peak at 0.45 ($p_{max}=0\%-n_{SNP}=2,447$) and decreased again until 0.33 ($p_{max}=45\%-n_{SNP}=11,707$) where it plateaued (Fig. 25d). Similarly, G_ASGM_A firstly decreased from 0.54 ($p_{max}=0\%-n_{SNP}=2,447$) to 0.46 ($p_{max}=5\%-n_{SNP}=5,620$) then progressively increased with the SNP dataset and plateaued at SNP

dataset $p_{max}=45\%-n_{SNP}=11,707$ with a prediction accuracy of 0.55 (Fig. 25d). Regarding the pedigree-based models, moderate prediction accuracies were obtained, i.e., 0.37 and 0.4, respectively for P_ASGM_A and P_PSAM_A.

Genomic prediction accuracies of NF ranged from 0.43 to 0.61 depending on the SNP dataset (Fig. 25e). Accuracies in both GS predictions depicted a high variation from an SNP dataset to another. For the pedigree-based model, prediction accuracies were 0.46 and 0.5, respectively for P_ASGM_A and P_PSAM_A (Fig. 25e).

On average over traits and SNP datasets, G_ASGM_A was more accurate (0.45) than G_PSAM_A (0.43), 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 0.13 (PF) to 0.59 (ABW) for G_PSAM_A. G_ASGM_A obtained a mean prediction accuracy greater than G_PSAM_A for five traits out of eight, with G_PSAM_A being on average more accurate than G_ASGM_A for AFW, NF and ABW (Table IX). 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 0.23 (PF) to 0.63 (ABW) for G_PSAM_A (Table IX), and G_ASGM_A was again more often better than G_PSAM_A (with G_PSAM_A being more accurate for PF, NF and ABW).

Considering the different SNP datasets and traits, G_ASGM_A gave higher prediction accuracy than G_PSAM_A in 58.3% of the cases, with the largest differences in prediction accuracy in favor of G_ASGM_A, up to 0.22 with OP at $p_{max} = 45\%-n_{SNP} = 11,707$ (although they were non-significant) (Fig. 24, Fig. 25 and Table X). Significant differences were only found in favor of G_PSAM_A, but they were scarce (i.e., only for NF in three SNP datasets, $p_{max}=5\%-n_{SNP}=5,620$, $p_{max}=10\%-n_{SNP}=6,898$ and $p_{max}=45\%-n_{SNP}=11,707$). Despite the overall lower prediction accuracies of G_PSAM_A compared to G_ASGM_A, G_PSAM_A was the most accurate method for ABW and NF with all the SNP datasets, except for NF with $p_{max}=75\%-n_{SNP}=15,054$. G_ASGM_A, therefore, appeared to be the best approach (i.e., generally more accurate, in addition to being easier to implement) for predicting clonal values for oil palm yield components, although G_PSAM_A could be worthwhile for some traits (ABW and NF here).

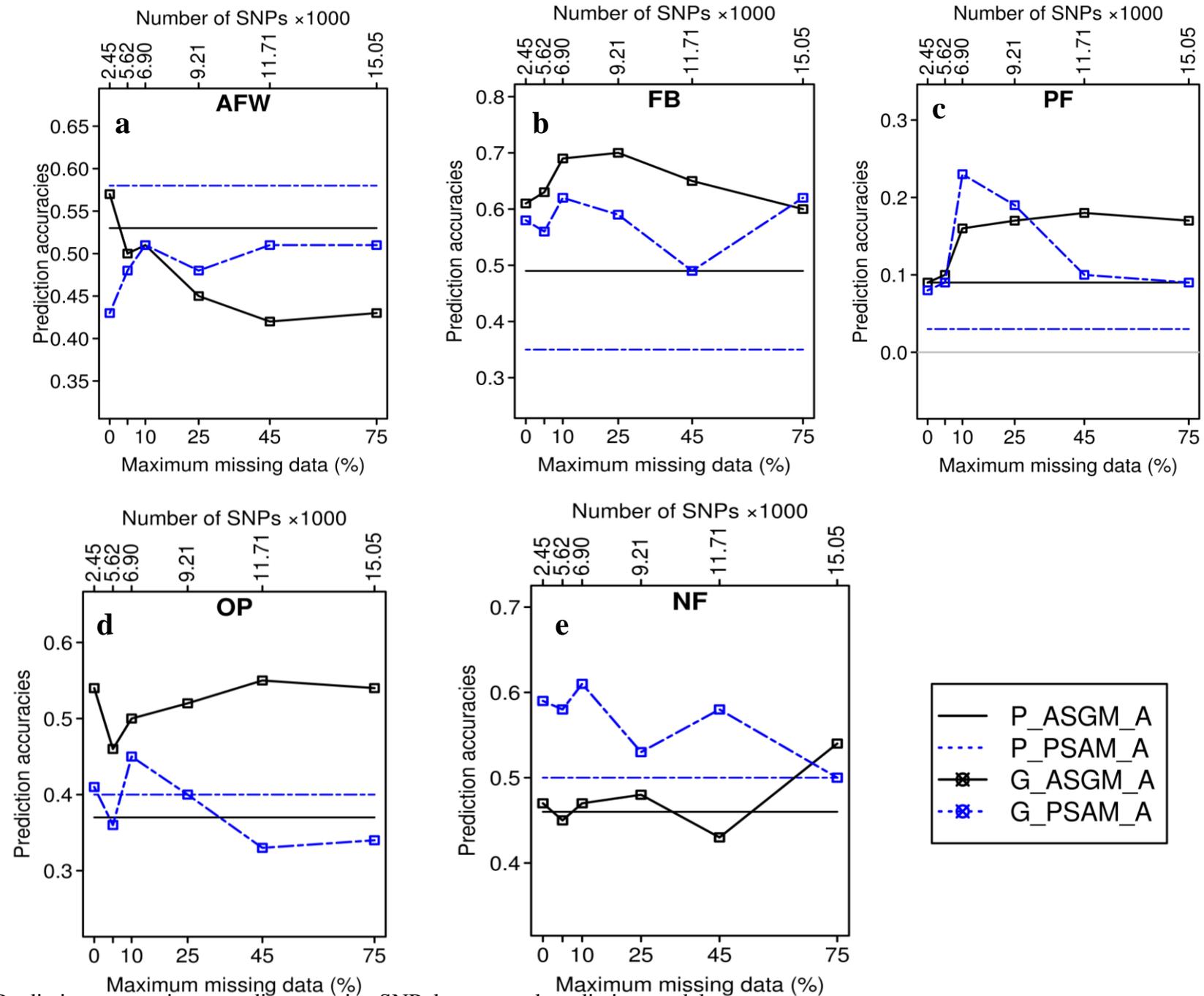


Fig. 25. Prediction accuracies according to traits, SNP datasets and prediction models. a: average fruit weight (AFW); b: fruit to bunch (FB) ratio; c: pulp to fruit (PF) ratio; d: oil to pulp (OP) ratio; e: number of fruits (NF) per bunch; pedigree-based prediction models: across-population SNP genotype models (P_ASGM_A), population-specific effects of SNP alleles models (P_PSAM_A); additive genomic prediction models: across-population SNP genotype models (G_ASGM_A), population-specific effects of SNP alleles models (G_PSAM_A).

Table IX. Mean prediction accuracies according to trait and prediction model.

Bunch production: bunch number (BN), average bunch weight (ABW) and total bunch production (FFB); bunch quality: 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); genomic prediction models: across-population SNP genotype models (G_ASGM_A), population-specific effects of SNP alleles models (G_PSAM_A). Values in brackets indicate the corresponding SNP dataset, defined on its maximum percentage of missing data.

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	0.49	0.57 (0%)	0.51 (10%/45%/75%)
FB	0.65	0.58	0.70 (25%)	0.62 (10%/75%)
PF	0.14	0.13	0.18 (45%)	0.23 (10%)
OP	0.52	0.38	0.55 (45%)	0.45 (10%)
NF	0.47	0.57	0.54 (75%)	0.61 (10%)
FFB	0.47	0.41	0.55 (10%)	0.51 (45%)
BN	0.31	0.30	0.37 (75%)	0.35 (75%)
ABW	0.53	0.59	0.58 (75%)	0.63 (5%)
Mean	0.45	0.43	0.51	0.49

Prediction accuracies could be broadly improved when relationship matrices were computed using SNPs (G_ASGM_A and G_PSAM_A) instead of genealogical data (control pedigree-based models P_ASGM_A and P_PSAM_A), in particular for three traits FB, BN and ABW. The maximum prediction accuracies of GS over all SNP datasets outperformed pedigree-based models for seven traits out of eight (except for AFW with G_PSAM_A) (Table XI, Fig. 24 and Fig. 25). The largest difference was observed in BN for $p_{max}=75\%-n_{snp}=15,054$, with G_ASGM_A accuracy being 0.67 higher than P_ASGM_A. Significant differences between GS models and their pedigree-based control models were found for five traits, with four traits (FB, OP, BN and ABW) where GS was the best and one trait (AFW) where pedigree-based models were more accurate (Table XI). The percentage of combinations of SNP datasets and traits where G_ASGM_A was more accurate than its control pedigree-based version reached 83.3%, against only 64.6% for G_PSAM_A.

Table X. 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: * $0.05 > P \geq 0.01$; ** $0.01 > P \geq 0.001$; *** $P < 0.001$.

Bunch production: bunch number (BN), average bunch weight (ABW) and total bunch production (FFB); bunch quality: 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); pedigree-based prediction models: across-population SNP genotype models (P_ASGM_A), population-specific effects of SNP alleles models (P_PSAM_A); genomic prediction models: across-population SNP genotype models (G_ASGM_A), population-specific effects of SNP alleles models (G_PSAM_A).

SNP dataset	Compared models	AFW	FB	PF	OP	NF	FFB	BN	ABW
	<i>P_ASGM_A</i> – <i>P_PSAM_A</i>	-0.06	0.15*	0.06	-0.03	-0.04	0.03	-0.25**	-0.04
0%-2,447	<i>G_ASGM_A</i> – <i>G_PSAM_A</i>	0.14	0.03	0.01	0.13	-0.12	0.05	-0.03	-0.12
5%-5,620	<i>G_ASGM_A</i> – <i>G_PSAM_A</i>	0.02	0.07	0.01	0.10	-0.13*	0.08	0.06	-0.11
10%-6,898	<i>G_ASGM_A</i> - <i>G</i> <i>PSAM_A</i>	0.00	0.07	-	0.05	-0.14*	0.08	-0.01	-0.08
				0.07					
25%-9,205	<i>G_ASGM_A</i> - <i>G</i> <i>PSAM_A</i>	-0.03	0.11	-	0.12	-0.05	0.13	0.00	-0.01
				0.02					
45%- 11,707	<i>G_ASGM_A</i> – <i>G_PSAM_A</i>	-0.09	0.16	0.08	0.22	-0.15*	-0.02	0.00	-0.03
75%- 15,054	<i>G_ASGM_A</i> - <i>G</i> <i>PSAM_A</i>	-0.08	-0.02	0.08	0.20	0.04	0.09	0.02	-0.01

The SNP dataset affected the prediction accuracy differently according to the trait and the model. 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 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 two traits (AFW and BN). For the other

traits, prediction accuracies remained stable or tended to decrease with increasing marker density and the maximum percentage of missing SNP data.

Table XI. 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: $*0.05 > P \geq 0.01$; $**0.01 > P \geq 0.001$; $***P < 0.001$.

Bunch production: bunch number (BN), average bunch weight (ABW) and total bunch production (FFB); bunch quality: 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); pedigree-based prediction models: across-population SNP genotype models (P_ASGM_A), population-specific effects of SNP alleles models (P_PSAM_A); genomic prediction models: across-population SNP genotype models (G_ASGM_A), population-specific effects of SNP alleles models (G_PSAM_A).

SNP dataset	Compared models	AFW	FB	PF	OP	NF	FFB	BN	ABW
0%-2,447	<i>P_ASGM_A</i> – <i>G_ASGM_A</i>	-0.04	-0.12	0.00	-0.17	-	0.07	-0.53**	-0.19
	<i>P_PSAM_A</i> – <i>G_PSAM_A</i>	0.15	-0.23*	-	-0.01	-	0.09	-0.32*	-0.26
5%-5,620	<i>P_ASGM_A</i> – <i>G_ASGM_A</i>	0.03	-0.14	-	-0.09	-	-	-0.56**	-0.28*
	<i>P_PSAM_A</i> – <i>G_PSAM_A</i>	0.10	-0.21	-	0.04	-	-	-0.25	-0.34**
10%-6,898	<i>P_ASGM_A</i> – <i>G_ASGM_A</i>	0.02	-0.20*	-	-0.13	-	-	-0.59**	-0.30*
	<i>P_PSAM_A</i> – <i>G_PSAM_A</i>	0.07	-0.27*	-	-0.05	-	-	-0.35*	-0.33*
25%-9,059	<i>P_ASGM_A</i> – <i>G_ASGM_A</i>	0.08	-0.20*	-	-0.15	-	-	-	-0.30**
	<i>P_PSAM_A</i> – <i>G_PSAM_A</i>	0.10	-0.24*	-	0.00	-	-	-0.39**	-0.27*
45%-11,425	<i>P_ASGM_A</i> – <i>G_ASGM_A</i>	0.11	-0.15	-	-0.18*	0.03	-	-	-0.30**
	<i>P_PSAM_A</i> – <i>G_PSAM_A</i>	0.07	-0.14	-	0.07	-	-	-0.38*	-0.29*
75%-15,054	<i>P_ASGM_A</i> – <i>G_ASGM_A</i>	0.10*	-0.11	-	-0.17	-	-	-	-
	<i>P_PSAM_A</i> – <i>G_PSAM_A</i>	0.07	-	-	0.06	0.00	-	-0.40*	-0.30*

However, the use of a different SNP dataset for each combination of trait and model seems unrealistic for the practical application of GS. Therefore, in order to identify the optimal SNP dataset(s) that would maximize GS accuracy, we computed for each GS prediction model and SNP dataset the mean prediction accuracy over the traits. For G_ASGM_A, this value increased with the SNP density (0.41 with 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 the subsequent SNP datasets. This shows that, for G_ASGM_A, the number of SNPs was of greater importance than the percentage of missing data per SNP. Mean prediction accuracy over the SNP datasets forming the plateau ranged from 0.17 (PF) to 0.66 (FB), and were close to the highest accuracies achieved over all the SNP datasets (Table IX). For G_ASGM_A, there was, therefore, a minimum of 6,898 SNPs required to reach maximum prediction accuracy on average over all traits. For G_PSAM_A, the results differed, with a peak in mean prediction accuracy at 0.47 with SNP dataset $p_{max}=10\%-n_{snp}=6,898$ and mean prediction accuracy decreasing when less SNPs were used, falling to 0.39 with $p_{max}=0\%-n_{snp}=2,447$, and decreasing when there were more missing data, falling to 0.41 with $p_{max}=75\%-n_{snp}=15,054$. This shows that G_PSAM_A was more sensitive to the SNP dataset than G_ASGM_A, making again G_PSAM_A less appealing. Therefore, for the final part of the study, we decided to focus on G_ASGM_A.

III.1.1.3. Comparison of prediction accuracies of PS and GS

Fig. 26 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 XII and Fig. 26). These two traits are known to have moderate to high heritability in the oil palm (Corley & Tinker, 2016) 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 XII).

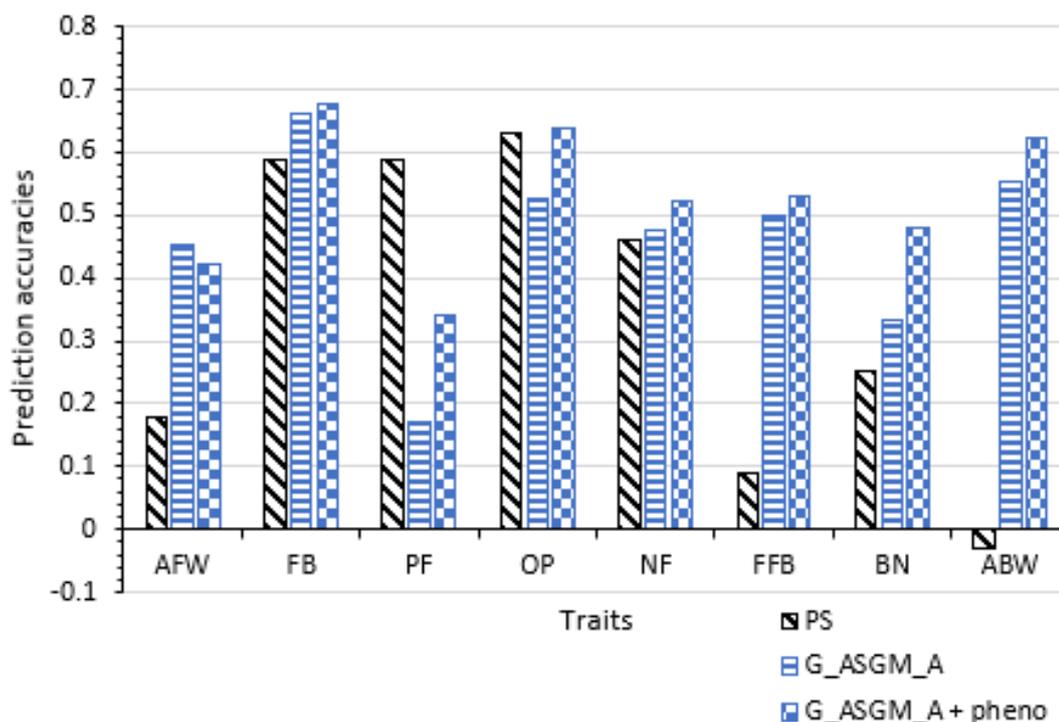


Fig. 26. Prediction accuracies on average over the best SNP datasets and according to the trait.

Prediction accuracies of phenotypic selection (PS); genomic prediction models: across-population SNP genotype models without phenotypic data (G_ASGM_A) and with phenotypic data (G_ASGM_A+pheno) of ortets, population-specific effects of SNP alleles models (G_PSAM_A). Bunch production: bunch number (BN), average bunch weight (ABW) and total bunch production (FFB); bunch quality: 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 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. 26). 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_ASGM_A+pheno for one trait (PF).

Table XII. Intensity and accuracy of phenotypic selection before clonal trials according to trait. Bunch production: bunch number (BN), average bunch weight (ABW) and total bunch production (FFB); bunch quality: average fruit weight (AFW), fruit to bunch (FB), pulp to fruit (PF), and oil to pulp (OP) ratios, and the number of fruits per bunch (NF).

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

III.1.2. Effect of the genotyping strategy to optimize prediction accuracy

III.1.2.1. Effect on prediction accuracy of using the genotyping strategy for the training population

In G_ASGM_Par models, prediction accuracies ranged from 0.15 for FB to 0.88 for AFW and in G_ASGM_Par+Hyb models, from 0.20 for FB to 0.88 for AFW (Fig. 27 and Fig. 28). Prediction accuracies for G_ASGM_Par+Hyb were higher than those of G_ASGM_Par for eight traits out of nine and the same for one trait (AFW), suggesting that training the model with genomic data of hybrid individuals in addition with those of parents increase prediction accuracies. Indeed, for AFW, prediction accuracies between G_ASGM_Par (0.88) and G_ASGM_Par+Hyb 0.88 were identical i.e., not significantly different (Fig. 27a). Prediction accuracies of FB was 0.15 for G_ASGM_Par and 0.2 for G_ASGM_Par+Hyb i.e., 25% higher, although without significant difference (Fig. 27b). Concerning PF, prediction accuracies were almost similar between (0.36) (Fig. 27c). Similarly, no significant difference was observed between G_ASGM_Par and G_ASGM_Par+Hyb for OP and NF (Fig. 27d, e). The only significant difference for bunch quality traits observed between G_ASGM_Par and G_ASGM_Par+Hyb was for OER. For this latter, the prediction accuracies of G_ASGM_Par was 0.52 while that of G_ASGM_Par+Hyb was 0.58 (Fig. 27f).

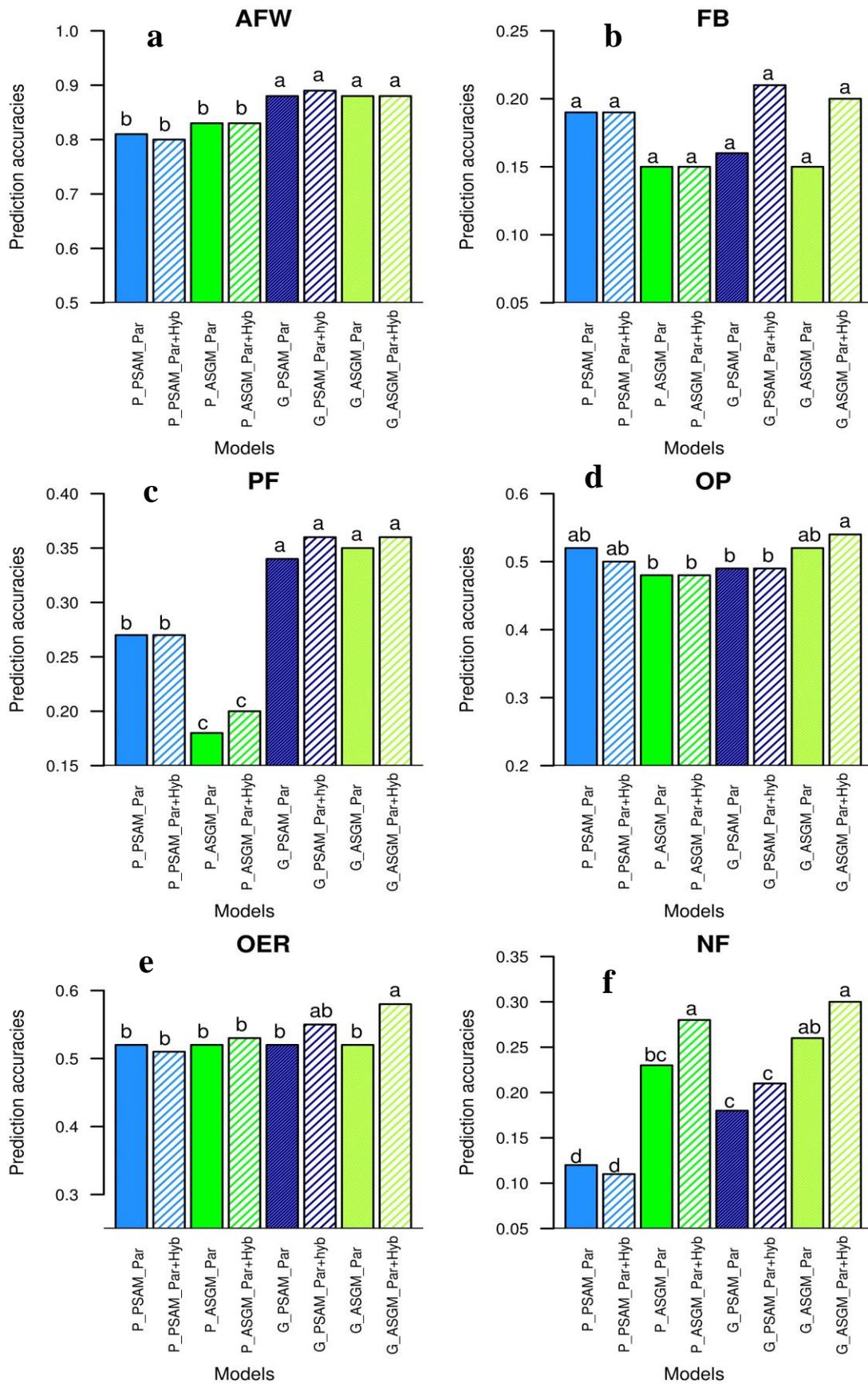


Fig. 27. Prediction accuracies of bunch quality traits according to prediction models. Values with the same letter are not significantly different within a trait at P = 5%.

a: average fruit weight (AFW); b: fruit to bunch (FB) ratio; c: pulp to fruit (PF) ratio; d: oil to pulp (OP) ratio; e: number of fruits per bunch (NF), f: oil extraction rate (OER). Pedigree-based prediction models: across-population SNP genotype models without hybrid individuals (P_ASGM_Par) and with hybrid individuals (P_ASGM_Par+Hyb), population-specific effects of SNP alleles models without hybrid individuals (P_PSAM_Par) and with hybrid individuals (P_PSAM_Par+Hyb); genomic prediction models: across-population SNP genotype models without hybrid individuals (G_ASGM_Par) and with hybrid individuals (G_ASGM_Par+Hyb), population-specific effects of SNP alleles models without hybrid individuals (G_PSAM_Par) and with hybrid individuals (G_PSAM_Par+Hyb). Genomic data of 399 hybrid individuals were used for bunch quality traits.

Prediction accuracies of G_ASGM_Par and G_ASGM_Par+Hyb were not significantly different for BN and FFB, respectively. The only significant difference of prediction accuracy between G_ASGM_Par and G_ASGM_Par+Hyb for bunch production traits was observed on ABW; with G_ASGM_Par+Hyb (0.7) being 4.3% more accurate than G_ASGM_Par (0.67) (Fig. 28).

Averaged over traits, G_ASGM_Par+Hyb had a prediction accuracy of 0.53, which was significantly higher than that of G_ASGM_Par, with 0.50 (Fig. 29), i.e., an increase of 6%. Among traits, the increase ranged from 0% for AFW to 33.3% for FB and it was significant for two traits, ABW (+4.5%) and OER (+11.5%) as aforementioned (Fig. 28 and Fig. 27).

In G_PSAM_Par, prediction accuracies ranged from 0.16 for FB to 0.88 for AFW and from 0.21 for both FB and NF to 0.89 for AFW in G_PSAM_Par+Hyb (Fig. 27 and Fig. 28). Eight traits out of nine had better prediction accuracies with G_PSAM_Par+Hyb than with G_PSAM_Par, and one trait (OP) had similar prediction accuracy. Like with the G_ASGM approach, prediction accuracy thus increased when hybrid molecular data were added to the molecular data of hybrid parents in the training set. In detail, the prediction accuracy of AFW for G_PSAM_Par+Hyb model (0.89) was slightly higher than that of G_PSAM_Par (0.88) (Fig. 27a). For FB, even though the prediction accuracy of G_PSAM_Par+Hyb model (0.21) was 25% higher than that of G_PSAM_Par (0.16), no significant difference was observed (Fig. 27b). The prediction accuracy of G_PSAM_Par+Hyb model for PF with 0.36 was slightly higher than that of G_PSAM_Par with 0.34 (Fig. 27c). The only trait whose prediction accuracies were identical between G_PSAM_Par+Hyb model and G_PSAM_Par was OP with 0.49 (Fig. 27d). Regarding OER and NF, prediction accuracies of G_PSAM_Par+Hyb (0.55 and 0.2, respectively) were higher than G_PSAM_Par (0.52 and 0.18) but not significant (Fig. 27e, f).

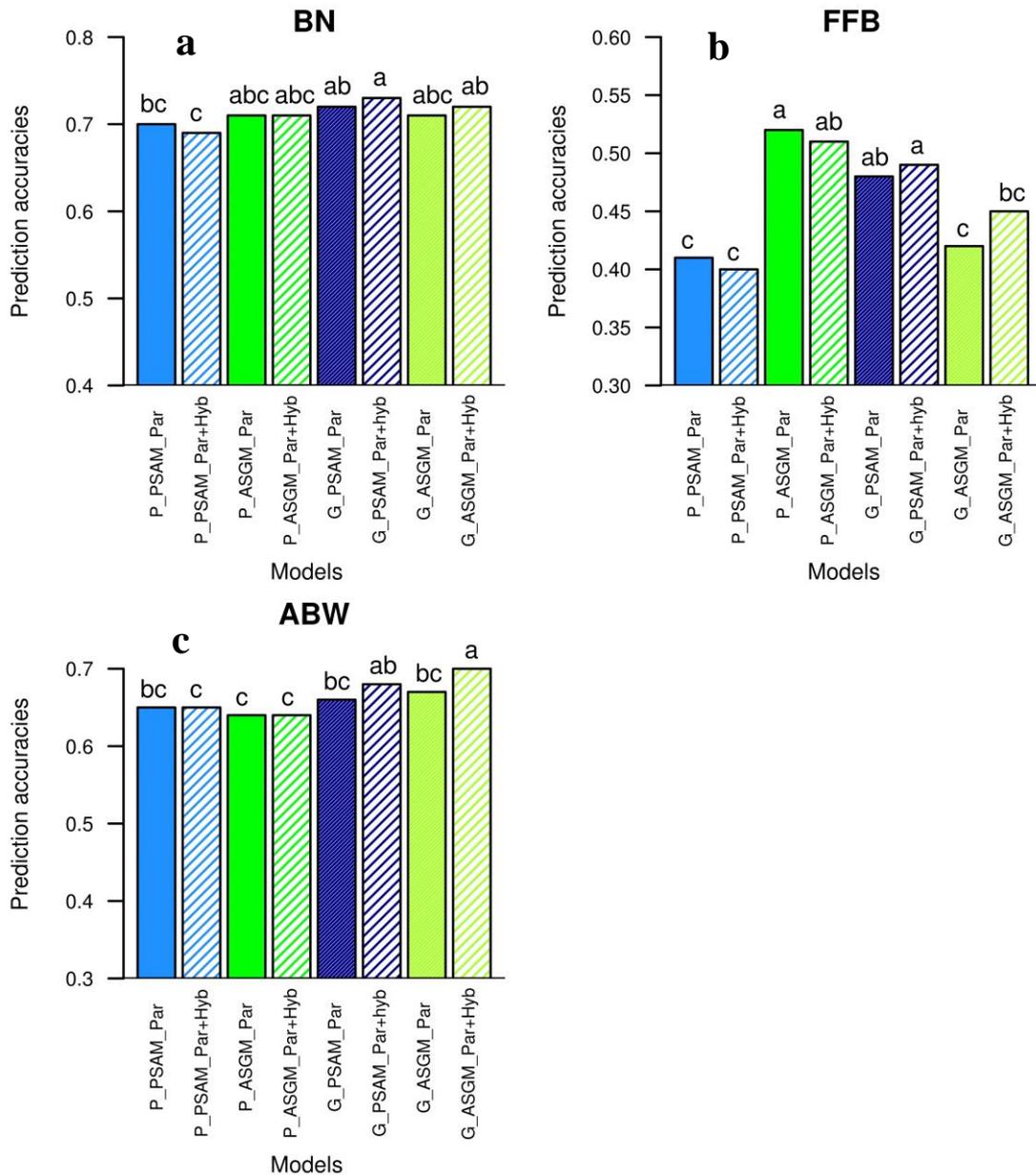


Fig. 28. Prediction accuracies of bunch production traits according to prediction models. Values with the same letter are not significantly different within a trait at $P = 5\%$.

Bunch production: bunch number (BN), average bunch weight (ABW) and total bunch production (FFB); pedigree-based prediction models: across-population SNP genotype models without hybrid individuals (P_ASGM_Par) and with hybrid individuals (P_ASGM_Par+Hyb), population-specific effects of SNP alleles models without hybrid individuals (P_PSAM_Par) and with hybrid individuals (P_PSAM_Par+Hyb); genomic prediction models: across-population SNP genotype models without hybrid individuals (G_ASGM_Par) and with hybrid individuals (G_ASGM_Par+Hyb), population-specific effects of SNP alleles models without hybrid individuals (G_PSAM_Par) and with hybrid individuals (G_PSAM_Par+Hyb). Genomic data of 397 hybrid individuals were used for bunch production traits.

For BN, the prediction accuracy of G_PSAM_Par+Hyb model (0.72) was slightly higher than that of G_PSAM_Par (0.71) (Fig. 28). Similarly, for FFB and ABW, prediction accuracies of G_PSAM_Par+Hyb (0.49 and 0.68, respectively) was slightly higher than those of G_PSAM_Par (0.48 and 0.66, respectively) (Fig. 28b, c).

The average prediction accuracy across traits of G_PSAM_Par+Hyb was 0.51, versus 0.49 for G_PSAM_Par (Fig. 29), i.e., an average increase of 4.1%, range: 0% for OP to 31.3% for FB (Fig. 27 and Fig. 28), although the increase was not statistically significant.

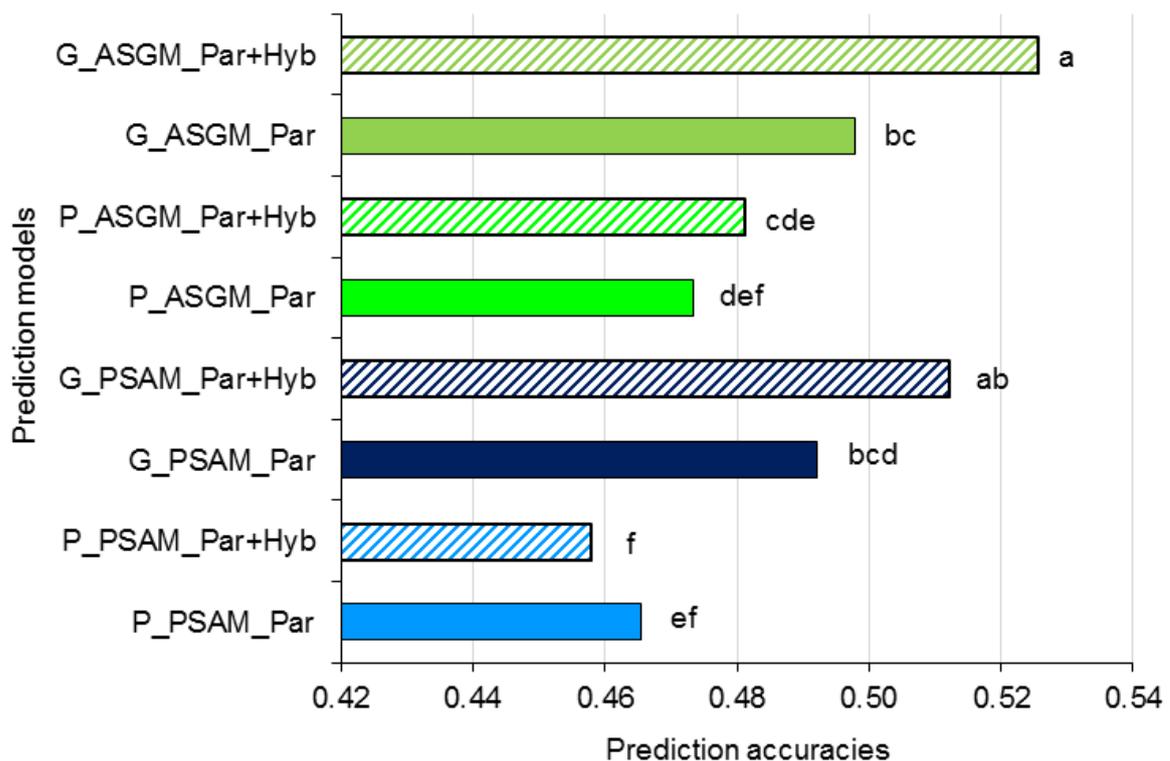


Fig. 29. Average prediction accuracies of prediction models across traits. Values with the same letter are not significantly different at $P = 5\%$.

Pedigree-based prediction models: across-population SNP genotype models without hybrid individuals (P_ASGM_Par) and with hybrid individuals (P_ASGM_Par+Hyb), population-specific effects of SNP alleles models without hybrid individuals (P_PSAM_Par) and with hybrid individuals (P_PSAM_Par+Hyb); genomic prediction models: across-population SNP genotype models without hybrid individuals (G_ASGM_Par) and with hybrid individuals (G_ASGM_Par+Hyb), population-specific effects of SNP alleles models without hybrid individuals (G_PSAM_Par) and with hybrid individuals (G_PSAM_Par+Hyb). Genomic data of 397 and 399 hybrid individuals were used for bunch production and bunch quality traits, respectively.

Regarding the control pedigree-based models, adding the genealogical data of hybrid individuals changed the prediction accuracies of traits in negligible and inconsistent ways (Fig. 27 and Fig. 28). The average prediction accuracy of G_PSAM_Par was 0.47 and 0.46 for

G_PSAM_Par+Hyb (+2.2%). On the other hand, the average prediction accuracies of G_ASGM_Par and G_ASGM_Par+Hyb were respectively 0.47 and 0.48 (-2.1%) (Fig. 27, Fig. 28 and Fig. 29).

III.1.2.2. Effect on prediction accuracy of the method used to model marker effects

Prediction accuracies were on average 2% higher with G_ASGM_Par than with G_PSAM_Par (Fig. 29). G_ASGM_Par was more accurate than G_PSAM_Par for four traits (ABW, PF, OP and NF, the difference being significant for NF, where the increase reached 44.4%) (Fig. 27 and Fig. 28). On the other hand, in comparison to G_PSAM_Par, G_ASGM_Par produced better prediction accuracies for three traits (BN, FFB and FB, with a significant increase for FFB (+14.3%)). Similar prediction accuracies were obtained for two traits (AFW and OER) (Fig. 27 and Fig. 28).

The prediction accuracy of G_ASGM_Par+Hyb was 4% higher than that of G_PSAM_Par+Hyb (Fig. 29). G_ASGM_Par+Hyb outperformed G_PSAM_Par+Hyb for four traits (ABW, OP, OER and NF), with significant differences for OP and NF, where prediction accuracy was, respectively, 10.2% and 43% higher than G_PSAM_Par+Hyb. G_ASGM_Par+Hyb underperformed G_PSAM_Par+Hyb for four traits (BN, FFB, FB and AFW), although the differences were never significant, and the prediction accuracies for one trait (PF) of the two modeling methods were similar (Fig. 27 and Fig. 28).

Concerning the pedigree-based models, the average prediction accuracies of P_PSAM_Par and P_ASGM_Par were similar, while P_ASGM_Par+Hyb was 4.3% more accurate than P_PSAM_Par+Hyb (Fig. 27, Fig. 28 and Fig. 29).

III.1.2.3. Comparison of GS models and control pedigree-based models

On average across traits, the prediction accuracy of GS models was 8.5% higher than that of the pedigree-based models, and the difference was always significant (Fig. 29). The average prediction accuracy of the pedigree-based models was 0.47 (range: 0.17 for FB to 0.82) for AFW (Fig. 27 and Fig. 28). The prediction accuracy of GS models was higher than that of pedigree-based models for eight traits and lower for one trait (FFB). The biggest difference was for PF, for which GS prediction accuracy was 52.2% higher (Table XIII).

Table XIII. Maximum prediction accuracies of traits.

Bunch production: bunch number (BN), average bunch weight (ABW) and total bunch production (FFB); bunch quality: average fruit weight (AFW), fruit to bunch (FB), pulp to fruit (PF), and oil to pulp (OP) ratios, number of fruits per bunch (NF), and oil extraction rate (OER); pedigree-based prediction models: across-population SNP genotype models without hybrid individuals (P_ASGM_Par) and with hybrid individuals (P_ASGM_Par+Hyb), population-specific effects of SNP alleles models without hybrid individuals (P_PSAM_Par) and with hybrid individuals (P_PSAM_Par+Hyb); genomic prediction models: across-population SNP genotype models without hybrid individuals (G_ASGM_Par) and with hybrid individuals (G_ASGM_Par+Hyb), population-specific effects of SNP alleles models without hybrid individuals (G_PSAM_Par) and with hybrid individuals (G_PSAM_Par+Hyb). Genomic data of 397 and 399 hybrid individuals were used for bunch production and bunch quality traits, respectively.

Traits	Best predictive models	Prediction accuracy
BN	G_PSAM_A_Par+Hyb	0.73
FFB	P_ASGM_A_Par	0.52
ABW	G_ASGM_Par+Hyb	0.70
AFW	G_PSAM_A_Par+Hyb	0.89
FB	G_PSAM_A_Par+Hyb	0.21
PF	G_PSAM_A_Par+Hyb/ G_ASGM_Par+Hyb	0.36
OP	G_ASGM_Par+Hyb	0.54
OER	G_ASGM_Par+Hyb	0.58
NF	G_ASGM_Par+Hyb	0.30

III.2. Discussion

The present thesis, evaluated empirically the benefit of using genomic data from $A \times B$ hybrid individuals for the genomic approach applied to oil palm (*Elaeis guineensis* Jacq.), using GS models and high throughput SNP genotyping (GBS). Two situations were considered: the evaluation of the efficiency of GS for clonal selection and the investigation of the effect of the genotyping strategy to optimize prediction accuracy. In both situations, the effect on prediction accuracy 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) were assessed.

III.2.1. Efficiency of genomic selection for clonal selection

III.2.1.1. 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 (Corley & Tinker, 2016). In our dataset, these traits indeed had high H^2 , with PS prediction accuracy >0.5 (Fig. 26) (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. 26), 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) (Soh *et al.*, 2017). 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 (Nyouma *et al.*, 2019).

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 detail as its efficiency also depends on the accuracy of the genomic estimated breeding values of the parents.

III.2.1.2. 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 (Cros *et al.*, 2018). However, G_PSAM_A usually did not perform better than G_ASGM_A, except for ABW and NF. We hypothesized that this was the consequence of stronger differences among Deli and La Mé populations at the QTLs controlling ABW and NF than at 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 ABW and BN, and noting that, although G_PSAM_A was not better than G_ASGM_A for BN, their results were actually very similar for this trait. This is in agreement with the results of Tisné *et al.* (2015), 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. The possibility for G_PSAM_A to outperform G_ASGM_A 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. 20b and a). 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 (Technow *et al.*, 2014).

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) (Zeng *et al.*, 2013). 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 G_PSAM_A. Our results corroborate those of Zeng *et al.* (2013) 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.

Ibáñez-Escriche *et al.* (2009) 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.* (2015) 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.* (2012) 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 (Singh *et al.*, 2013b), the investigated range of SNP density was similar, going from 0.8 to 8.4 SNP per Mbp.

Moreover, Lopes *et al.* (2017) obtained similar prediction accuracies between G_ASGM_A and G_PSAM_A with high SNP density (31,930 SNPs). In our study, the only SNP dataset where G_PSAM_A outperformed G_ASGM_A on average on all traits was a dataset with intermediate number of SNPs and intermediate percentage of missing data per SNP, $p_{max}=10\%-n_{SNP}=6,898$, with mean G_PSAM_A prediction accuracy of 0.47 against 0.46 for G_ASGM_A. This result, therefore, differs from those of Technow *et al.* (2012) and Lopes *et al.* (2017), likely as a consequence of the fact that, in our study, SNP density varied with SNP quality, with higher SNP numbers meaning a higher percentage of missing data. This indicates that the SNP dataset must be chosen carefully before applying G_PSAM A. From this point of view, G_ASGM_A appeared advantageous, as its mean accuracy over the traits remained at its maximum once sufficient SNP density was reached, regardless of the percentage of missing data. The fact that for G_ASGM_A the number of SNPs was of greater importance than the percentage of missing data per SNP indicates that Beagle 4.0 efficiently imputed the missing data. Therefore, the existence of an optimal SNP dataset for G_PSAM_A suggests that phasing errors increase with the percentage of missing data per SNP and when decreasing the marker density.

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 (Ong-Abdullah *et al.*, 2015)), 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 family effects, while pedigree-based models can only account, at best, for family effects, as already found in previous oil palm GS studies (Nyouma *et al.*, 2019).

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 one trait (AFW), while that never happened with G_ASGM_A. Also, the overall inferiority of G_PSAM_A to G_ASGM_A occurred while P_PSAM_A was actually better than P_ASGM_A for five traits out of eight. This looks contradictory and suggests that the performance of G_PSAM_A could have been reduced by phasing errors as aforementioned. Also, many studies comparing G_ASGM_A and G_PSAM_A were carried out by simulation with known phases (Technow *et al.*, 2012; Zeng *et al.*, 2013; Esfandyari *et al.*, 2015), and therefore possible phasing errors in our study could also be the cause of the discrepancies observed between our results and the results obtained in simulation studies. Investigating other phasing approaches seems therefore of interest in the oil palm context.

III.2.1.3. Genotyped individuals for training

In this study, to make GS predictions more cost-effective, the genotypes of the phenotyped hybrid individuals constituting the training set were reconstructed using the molecular data of their parents, with G_ASGM, or not used in the model, with G_PSAM. Both modeling approaches, therefore, assume that the mean genotype in a hybrid family (i.e., the mean number of copies of the minor allele over the individuals making the family) expected from the parental genotypes is the same as the actual mean genotype. Nevertheless, in the case

of allele segregation distortion at a locus, the mean genotype in a hybrid family would significantly deviate from the mean genotype expected from the parental genotypes, and this could reduce the GS accuracy. Indeed, high numbers of distorted markers can be found in plants: Li *et al.* (2015) and Zuo *et al.* (2019) found more than 10% of markers (SNPs and SSRs) significantly distorted. For future studies, it would be of great interest to compare the approach used here with predictions made using real hybrid genotypes, and to measure the differences in terms of GS accuracy and cost.

III.2.1.4. Prediction of dominance effects

GS prediction accuracies were not significantly enhanced by adding dominance effects. Including dominance effects in the statistical model sometimes slightly increased or reduced accuracies, depending on the traits and the SNP datasets, revealing a negligible genetic dominance variance captured by the model compared to the total genetic variance, as already observed with genomic predictions for performances of oil palm hybrid crosses (Cros *et al.*, 2017). We assume this was a consequence of reciprocal recurrent selection, which generated the contrasted allele frequencies we observed across Deli and La Mé populations, thus decreasing the ratio of SCA variance to GCA variance (Reif *et al.*, 2007) and making dominance effects absorbed by the GCAs or the population mean (Technow *et al.*, 2014).

III.2.2. Effect of the genotyping strategy to optimize prediction accuracy

III.2.2.1. Using genomic data of hybrid individuals to train the GS model

Models including genomic data on hybrids performed better than the corresponding parental models, or at least produced equivalent results, because, at SNPs that are heterozygotes in at least one parent, the genomic information on the hybrid individuals captures the segregation of the parental alleles within the hybrid crosses, and also accounts for possible segregation distortion. The superiority of models that included genomic data on hybrids was demonstrated for the two types of models tested here, ASGM and PSAM, underlining the robustness of the approach.

However, the prediction accuracies were only slightly increased compared to when only using the parental genomic data, or even similar for some traits. This was probably due to the low number of genotyped hybrid individuals. Indeed, only 2.66% and 1.76% hybrid individuals were genotyped in the calibration set for bunch production and quality traits, respectively, the

genotypes of the remaining >97% hybrid individuals being replaced by the average genotypes expected from the cross of their two parents. Cros *et al.* (2015a) in a simulation study found that, although genotyping 300 training hybrid individuals (i.e., 25% less than what we used here) led to lower genetic progress than using only the parental genotypes, genetic progress increased with the number of hybrid individuals genotyped and, with 1,000 and 1,700 genotyped training hybrid individuals, reached much greater values than using only the parental genotypes. Thus, in the case of oil palm, 400 seems the minimum number of training hybrid individuals to genotype. Here we could not investigate how GS prediction accuracy was affected by an increase in the number of genotyped hybrids. In lodgepole pine, Ukrainetz & Mansfield (2020) considering a population of 1,569 trees, found that GS prediction accuracy increased little with more than 40% of the training trees genotyped. In oil palm, this was so far only investigated by a simulation study (Cros *et al.*, 2015a) which showed that genotyping 1,700 hybrid individuals only slightly improved the results compared to when genotyping 1,000 individuals. An empirical study is lacking in oil palm on this aspect.

As mentioned above, genotyping hybrid individuals allows taking advantage of the segregation of the parental alleles within the hybrid crosses. The magnitude of this segregation is directly affected by the heterozygosity of the parents. In the current study, the percentage of heterozygote SNPs was low, under the effect of generations where inbreeding was commonly used, by selfing or by mating related selected individuals. The percentage of heterozygote SNPs was thus on average 6.6%, ranging from 3.1% to 11.2%, for the parents of group A and 8.1%, ranging from 3.3% to 14.1%, for the parents of group B. Therefore, it is worth genotyping training hybrid individuals even with a low percentage of heterozygosity in parents.

The models used here that did not include genomic information of hybrid individuals assumed that the genotypes among individuals of a given hybrid cross derived from the parental genotypes following Mendelian rules. However, segregation distortion, i.e. the deviation between the expected Mendelian allele frequencies and the actual allele frequency, is a common phenomenon in animal and plant reproduction (Lyttle, 1991; Taylor & Ingvarsson, 2003; Diouf & Mergeai, 2012). It is mainly caused by zygotic and gametic selection (pollen abortion, pollen tube competition and competitive fertilization) (Lyttle, 1993; Xian-Liang *et al.*, 2006; Xu *et al.*, 2013). In *E. guineensis*, it has been reported in several mapping studies. For example, Ting *et al.* (2014) found that 9.4% of SSR markers and 7.9% of SNPs showed segregation distortion at $P < 5\%$ in a Deli \times Yangambi hybrid cross, and Gan *et al.* (2018) found consistent results, i.e. 9.6% and 11% of markers with segregation distortion at $P < 5\%$ in two crosses of Binga \times

Yangambi-AVROS origin, among a set of SSR, DArT and SNP markers. Models including genomic information of hybrid individuals in the training dataset take into account the alleles distortion segregation (although partially, as only a sample of the hybrid individuals are genotype), and this is another advantage of these models. However, the data available here did not allow making a distinction between the effect of capturing within crosses genetic variability and the effect of taking into account segregation distortion. This could be further investigated using a population with larger full-sib families.

The drawback of genotyping hybrid individuals to train the GS model is that it increases costs and that it leads to GS statistical analyses that require extensive computer resources and become time-consuming. Genotyping only a sample of the phenotyped hybrids appears relevant, but further studies should investigate the optimal number of hybrid individuals to genotype to optimize the genetic progress per unit cost.

III.2.2.2. Effect of modelling of markers on prediction accuracy

The differences between G_ASGM and G_PSAM were usually small (on average 3%, with the exceptions of NF and FFB, with differences reaching 40% and 11.4%, respectively), which is in agreement with previous studies (Ibáñez-Escriche *et al.*, 2009; Technow *et al.*, 2012). Although G_ASGM was on average better than G_PSAM, the best method differed according to traits. Thus, for NF G_ASGM was 40% more accurate than G_PSAM while for FFB G_PSAM was 11.4% more accurate than G_ASGM. This is likely related to differences in the level of genetic divergence between the heterotic groups, A and B at the genes controlling the traits. Technow *et al.* (2012) indeed indicated that PSAM is most beneficial under low persistence of phases among parental populations, implying that the relative performance of G_PSAM and G_ASGM is affected by marker density and by the history of the parental populations, e.g., the number of generations since divergence. This aspect requires further investigation in oil palm. Another possible explanation of the mean superiority of G_ASGM over G_PSAM would be that the dataset used here did not allow taking full advantage of the PSAM approach. Indeed, PSAM is more challenging to implement, as it is more complex, with more variances and effects to estimate, and therefore requires a larger training population than ASGM. The performance of G_PSAM in oil palm might therefore increase using more hybrid individuals with phenotypic and genomic data. In addition, G_PSAM is affected by phasing errors. Using SNP array genotyping instead of GBS could make G_PSAM more efficient, as the lower percentage of missing data and genotyping errors with SNP arrays would improve

phasing. Also, other phasing approaches could be investigated. Here, in a preliminary analysis (not shown), the AlphaImpute software (AI) (Hickey *et al.*, 2012; Antolín *et al.*, 2017), which was used with pig crossbred data in Lopes *et al.* (2017), was tested for imputation and phasing, but it resulted in lower accuracies than Beagle 4.0 (Browning & Browning, 2007), for both GS modelling approaches.

Despite these possibilities for improving G_PSAM, the superiority of G_ASGM was already noted in a previous oil palm study, where a training population comprising the present Deli × La Mé crosses was used to predict the clonal values of hybrid individuals (Nyouma *et al.*, 2020). However, the current study extended the previous conclusion: indeed, in Nyouma *et al.* (2020), G_PSAM_Par performed better than G_ASGM_Par in different traits, i.e., ABW and NF against BN, FFB and FB here, which questions the robustness of the PSAM approach. As Nyouma *et al.* (2020) concluded that the G_ASGM approach should be preferred in oil palm as it was on average slightly more accurate, less sensitive to SNP dataset (i.e., SNP density and percentage of missing data) and easier to implement than PSAM, the comparison of the results of the two studies, therefore, adds a new element in favor of the use of G_ASGM in oil palm (Nyouma *et al.*, 2020).

GS models were usually more accurate than their corresponding pedigree-based control models. This confirmed that GS predictions can account for individual genetic effects (Mendelian sampling terms) and family genetic effects in the parental populations, while pedigree-based models can only account, at best, for family effects. We also noted that on average over PSAM and ASGM, adding genealogical information on the hybrid individuals did not change the prediction accuracies of the pedigree-based prediction models. This was expected as only the genomic information of the hybrid individuals can bring extra information to the model in terms of relationships, as the pedigree attributes the same relationship to all the full-sib hybrid individuals.

CHAPTER IV. CONCLUSION, PERSPECTIVES AND RECOMMENDATIONS

IV.1. Conclusion

Genomic selection (GS) is a major asset for the genetic improvement of crude palm oil yield in oil palm (*Elaeis guineensis* Jacq.) in order to supply the increasing world demand. For that purpose, the current study aimed at empirically evaluating the interest of using genomic data from A × B hybrid individuals for the genomic approach applied to oil palm. It appeared that GS can contribute to palm oil yield increase through clonal selection or parent selection for hybrid creation.

The evaluation of the efficiency of GS for clonal selection showed that clonal selection of oil palm can largely be improved thanks to the genomic approach. Indeed, GS prediction accuracies for ortets without phenotypic data records extended from 0.08 to 0.7 according to the trait, GS model and SNP dataset. The G_ASGM_A approach was better for predicting clonal values than G_PSAM_A, as it was on average slightly more accurate, less sensitive to the SNP dataset (i.e., SNP density and percentage of missing data) and easier to implement. However, G_PSAM_A appeared interesting for ABW and NF traits. The G_ASGM_A model required at least 7,000 SNPs to perform best, with the percentage of missing data per SNP being of secondary importance. In these conditions, G_ASGM_A gave higher prediction accuracies than current phenotypic selection for six traits out of eight. The annual genetic progress of clonal oil palm breeding for yield can be increased by replacing the current phenotypic ortet preselection before clonal trials either by genomic ortet preselection on most of the yield components among a large population of the best possible crosses (produced based on the results of the progeny tests) at the juvenile stage or by ortet preselection at the mature stage on all the yield components using jointly the genomic and phenotypic data of the ortet selection candidates.

Our findings on the evaluation of the effect of two strategies to optimize the GS accuracy indicated that, despite the relatively small number of hybrid individuals genotyped and the low level of heterozygosity in the parents, prediction accuracies were in most cases improved (or, at least, similar) when genomic information of hybrid individuals were added to the training dataset, compared to when using only the parental genomic information. The best GS approach investigated here, i.e., with the ASGM model and genotyping around 400 hybrid individuals, reached a mean prediction accuracy over traits of 0.53.

Moreover, the ASGM approach, i.e., using a model that does not take into account the parental origin of the marker alleles, is recommended for oil palm data, as it gives higher prediction accuracies on average over traits, performs best on more traits and is more robust over populations and SNP datasets than the PSAM approach, with population-specific marker allele effect.

IV.2. Perspectives

The current work showed the potential of GS for the genetic improvement of palm oil yield components. However, in order to meet world demand while simultaneously minimizing environmental impacts, future researches should focus on:

- the evaluation of different phasing approaches than Beagle;
- optimizing the prediction accuracies for all traits;
- optimizing the training population;
- optimizing the prediction model;
- the evaluation of the use of multi-omics data (transcriptomics, proteomics, etc.) for the training;
- evaluation of the effect of modeling of $G \times E$ interactions on prediction accuracy;
- the identification of the optimal number of hybrid individuals to genotype in order to maximize the selection response per unit cost, and better understand the factors controlling the relative performance of ASGM and PSAM approaches in hybrid crops.

IV.3. Recommendations

In order to increase the genetic gain in oil palm, it is recommended to oil palm breeding programs:

- to perform a preselection of ortet clones at the mature stage on all the yield components jointly using ortet genotypes and phenotypes;
- to make genomic preselection of ortet clones on all the yield components, among a large population of the best possible crosses at nursery stage;
- to utilize the across-population SNP genotype models (ASGM) for genomic prediction in oil palm yield components;
- to train genomic models using genomic data of the hybrid parents plus a sample of hybrid individuals.

REFERENCES

- Anonymous, 2010. *Global forest resources assessment 2010: Main report*. Food & Agriculture Org, Rome 376 p.
- Anonymous, 2013. Identity by descent. Wikipedia. Accessed September 1, 2020.
- Anonymous, 2020a. <https://www.palmelit.com/en/about-us>. Accessed November 09, 2020.
- Anonymous, 2020b. <http://www.fao.org/faostat/en/#data/QC/visualize>. Accessed November 09, 2020.
- Anonymous, 2020c. <http://www.fas.usda.gov/data/oilseeds-world-markets-and-trade>. Accessed January 13, 2020.
- Antolín R., Nettelblad C., Gorjanc G., Money D. & Hickey J. M., 2017. A hybrid method for the imputation of genomic data in livestock populations. *Genet. Sel. Evol.*, 49 (30): 1–17.
- Arolu I. W., Raffii M. Y., Marjuni M., Hanafi M. M., Sulaiman Z., Rahim H. A., Kolapo O. K., Abidin M. I. Z., Amiruddin M. D., Kushairi Din A. & Nookiah R., 2016. Genetic variability analysis and selection of pisifera palms for commercial production of high yielding and dwarf oil palm planting materials. *Ind. Crops Prod.*, 90: 135–141.
- Baird N. A., Etter P. D., Atwood T. S., Currey M. C., Shiver A. L., Lewis Z. A., Selker E. U., Cresko W. A. & Johnson E. A., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD Markers. *PLoS One*, 3 (10): 1–7.
- Baker W. J., Norup M. V., Clarkson J. J., Couvreur T. L., Dowe J. L., Lewis C. E., Pintaud J.-C., Savolainen V., Wilmot T. & Chase M. W., 2011. Phylogenetic relationships among Arecoid palms (Arecaceae: Arecoideae). *Ann. Bot.*, 108 (8): 1417–1432.
- Baumung R., Sölkner J. & Essl A., 1997. Correlation between purebred and crossbred performance under a two-locus model with additive by additive interaction. *J. Anim. Breed. Genet.*, 114 (1–6): 89–98.
- Beirnaert A. & Vanderweyen R., 1941. Contribution à l'étude génétique et biométrique des variétés d'*Elaeis guineensis* Jacq. *Publ. Inst. Nat. Etude Agron. Congo Belge Ser. Sci.* 27: 1–101.
- Billotte N., Jourjon M., Marseillac N., Berger A., Flori A., Asmady H., Adon B., Singh R., Nouy B. & Potier F., 2010. QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis* Jacq.). *Theor. Appl. Genet.*, 120 (8): 1673–1687.
- Boucher W., 1988. Calculation of the inbreeding coefficient. *J. Math. Biol.*, 26 (1): 57–64.
- Bouvet J.-M., Makouanzi G., Cros D. & Vigneron P., 2016. Modeling additive and non-additive effects in a hybrid population using genome-wide genotyping: prediction accuracy implications. *Heredity*, 116 (2): 146–157.
- Breure C. & Bos I., 1992. Development of elite families in oil palm (*Elaeis guineensis* Jacq.). *Euphytica*. 64 (1): 99–112.
- Breure C. J. & Verdooren L. R., 1995. Guidelines for testing and selecting parent palms in oil palm, practical aspects and statistical methods. *ASD Oil Palm Pap.* 9: 1-68.

- Browning S. R. & Browning B. L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, 81 (5): 1084–1097.
- Butler R. A., Koh L. P. & Ghazoul J., 2009. REDD in the red: palm oil could undermine carbon payment schemes. *Conserv. Lett.*, 2 (2): 67–73.
- De Los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D. & Calus M. P., 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193 (2): 327–345.
- Cappa E. P., de Lima B. M., da Silva-Junior O. B., Garcia C. C., Mansfield S. D. & Grattapaglia D., 2019. Improving genomic prediction of growth and wood traits in Eucalyptus using phenotypes from non-genotyped trees by single-step GBLUP. *Plant Sci.* 284: 9–15.
- Carvalho A. D. F. de, Fritsche Neto R. & Geraldi I. O., 2008. Estimation and prediction of parameters and breeding values in soybean using REML/BLUP and Least Squares. *Crop Breed. Appl. Biotechnol.*, 8 (3): 219–224.
- Chevalier A., 1943. Taxonomie, biogéographie et sélection des Palmiers du genre *Elaeis*. *J. Agric. Tradit. Bot. Appliquée*. 23 (266): 295–307.
- Cheverud J. M. & Routman E. J., 1995. Epistasis and its contribution to genetic variance components. *Genetics*, 139 (3): 1455–1461.
- Christensen O. F., Madsen P., Nielsen B., Ostersen T. & Su G., 2012. Single-step methods for genomic evaluation in pigs. *Animal*, 6 (10): 1565–1571.
- Christensen O., Madsen P., Nielsen B. & Su G., 2014. Genomic evaluation of both purebred and crossbred performances. *Genet. Sel. Evol.*, 46 (1): 1–9.
- Clark S. A. & van der Werf J., 2013. Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. In: Gondro C., van der Werf J. & Hayes B. (eds). *Genome-Wide Association Studies and Genomic Prediction*. Springer, New York-Heidelberg-Dordrecht-London: 321–330.
- Cochard B., 2008. *Etude de la diversité génétique et du déséquilibre de liaison au sein de populations améliorées de palmier à huile (Elaeis guineensis Jacq.)*. PhD thesis, Montpellier SupAgro, 272 p.
- Cochard B., Adon B., Kouame R. K., Durand-Gasselín T. & Amblard P., 2001. Intérêts des semences commerciales améliorées de palmier à huile (*Elaeis guineensis* Jacq.). *Ol. Corps Gras Lipides*. 8 (6): 654–658.
- Cochard B., Durand-Gasselín T. & PalmElit S., 2018. Advances in conventional breeding techniques for oil palm. In: Rival A. (ed). *Achieving sustainable cultivation of oil palm*. Burleigh Dodds Science Publishing, Cambridge: 133–160.
- Cockerham C. C., 1954. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, 39 (6): 859.
- Comstock R. E., Robinson H. F. & Harvey P. H., 1949. A breeding procedure designed to make maximum use of both general and specific combining ability. *Agron. J.*, 41 (8): 360–367.
- Conner J. K. & Hartl D. L., 2004. *A primer of ecological genetics*. Sinauer Associates Incorporated, Sunderland, 304 p.

- Corley R. H. V., 2006. Potential yield of oil palm—an update. *In: International Society of Oil Palm Breeders: Symposium on Yield Potential in the Oil Palm*. Phuket, 27–28 Nov 2006.
- Corley R. H. V., 2009. How much palm oil do we need? *Environ. Sci. Policy*, 12 (2): 134–139.
- Corley R. H. V. & Tinker P. B., 2016. *The oil palm*. John Wiley & Sons, Ltd, Chichester-West Sussex, 639 p.
- Corley R. & Law I., 1997. The future for oil palm clones. *In: Proc Int. Planters Conf. Incorpor. Soc. Kuala Lumpur*. 279–289.
- Covarrubias-Pazarán G., 2016. Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One*. 11 (6):1-15.
- Cronquist A. & Takhtadzhian A. L., 1981. *An integrated system of classification of flowering plants*. Columbia University Press, New York, 1262 p.
- Cros D., 2014. *Etude des facteurs contrôlant l'efficacité de la sélection génomique chez le palmier à huile (Elaeis guineensis Jacq.)*. PhD thesis, Montpellier SupAgro, 204 p.
- Cros D., Bocs S., Riou V., Ortega-Abboud E., Tisné S., Argout X., Pomiès V., Nodichao L., Lubis Z. & Cochard B., 2017. Genomic preselection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses. *BMC Genomics*, 18 (1): 839.
- Cros D., Denis M., Bouvet J.-M. & Sánchez L., 2015a. Long-term genomic selection for heterosis without dominance in multiplicative traits: case study of bunch production in oil palm. *BMC Genomics*, 16 (1): 651.
- Cros D., Denis M., Sánchez L., Cochard B., Flori A., Durand-Gasselín T., Nouy B., Omoré A., Pomiès V., Riou V., Suryana E. & Bouvet J.-M., 2015b. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). *Theor. Appl. Genet.*, 128 (3): 397–410.
- Cros D., Mbo-Nkoulou L., Bell J. M., Oum J., Masson A., Soumahoro M., Tran D. M., Achour Z., Le Guen V. & Clément-Demange A., 2019. Within-family genomic selection in rubber tree (*Hevea brasiliensis*) increases genetic gain for rubber production. *Ind. Crops Prod.* 138: 1-13.
- Cros D., Sánchez L., Cochard B., Samper P., Denis M., Bouvet J.-M. & Fernández J., 2014. Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population. *Theor. Appl. Genet.*, 127 (4): 981–994.
- Cros D., Tchounke B. & Nkague-Nkamba L., 2018. Training genomic selection models across several breeding cycles increases genetic gain in oil palm in silico study. *Mol. Breed.*, 38 (7): 89–101.
- Danecek P., Auton A., Abecasis G., Albers C. A., Banks E., DePristo M. A., Handsaker R. E., Lunter G., Marth G. T. & Sherry S. T., 2011. The variant call format and VCFtools. *Bioinformatics*, 27 (15): 2156–2158.
- De Los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel K. & Cotes J. M., 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182 (1): 375–385.
- De Souza C., 1992. Interpopulation genetic variances and hybrid breeding programs. *Rev. Bras. Genet.*, 15: 643–656.

- Demol J., 2002. *Amélioration des plantes: application aux principales espèces cultivées en régions tropicales*. Presses Agronomiques de Gembloux, 581 p.
- Dempster A. P., Laird N. M. & Rubin D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Methodol.*, 39 (1): 1–22.
- Diouf F. B. H. & Mergeai G., 2012. Distorsions de ségrégation et amélioration génétique des plantes (synthèse bibliographique). *BASE*, 16 (4): 499–508.
- Doolittle D. P., 1987. *Population genetics: basic principles*. Springer Science & Business Media, Berlin-Heidelberg, 264 p.
- Dransfield J., Uhl N. W., Asmussen C. B., Baker W. J., Harley M. M. & Lewis C. E., 2005. A new phylogenetic classification of the palm family, Arecaceae. *Kew Bull.*, 559–569.
- Durán R., Isik F., Zapata-Valenzuela J., Balocchi C. & Valenzuela S., 2017. Genomic predictions of breeding values in a cloned *Eucalyptus globulus* population in Chile. *Tree Genet. Genomes.*, 13 (4): 1-12.
- Durand-Gasselín T., Kouame R. K., Cochard B., Adon B. & Amblard P., 2000. Diffusion variétale du palmier à huile (*Elaeis guineensis* Jacq.). *Ol. Corps Gras Lipides.*, 7 (2): 207–214.
- Elshire R. J., Glaubitz J. C., Sun Q., Poland J. A., Kawamoto K., Buckler E. S. & Mitchell S. E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, 6 (5): e19379.
- Esfandyari H., Sørensen A. C. & Bijma P., 2015. A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genet. Sel. Evol.*, 47 (1): 64-76.
- Falconer D. & Mackay T., 1996. *Introduction to quantitative genetics*. Longman, Harlow-Essex, 464 p.
- Fisher R. A., 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Philos. Trans. R. Soc. Edinb.*, 52: 399–433.
- Fitzherbert E. B., Struebig M. J., Morel A., Danielsen F., Brühl C. A., Donald P. F. & Phalan B., 2008. How will oil palm expansion affect biodiversity? *Trends Ecol. Evol.*, 23 (10): 538–545.
- Friedrich S., Konietzschke F. & Pauly M., 2017. GFD: An R Package for the Analysis of General Factorial Designs. *J. Stat. Softw. Code Snippets.*, 79 (1): 1–18.
- Gallais A., 2011. *Méthodes de création de variétés en amélioration des plantes*. Quae, Versailles, 278 p.
- Gan S. T., Wong W. C., Wong C. K., Soh A. C., Kilian A., Low E.-T. L., Massawe F. & Mayes S., 2018. High density SNP and DArT-based genetic linkage maps of two closely related oil palm populations. *J. Appl. Genet.*, 59 (1): 23–34.
- Gascon J. P. & Berchoux C., 1964. Caractéristique de la production d'*Elaeis guineensis* Jacq. de diverses origines et de leurs croisements - Application à la sélection du palmier à huile. *Oléagineux*, 19 (2): 75–84.
- Gengler N., Misztal I., Bertrand J. & Culbertson M., 1998. Estimation of the dominance variance for postweaning gain in the US Limousin population. *J. Anim. Sci.*, 76 (10): 2515–2520.
- Gilmour A. R., Thompson R. & Cullis B. R., 1995. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 1440–1450.

- Glaubitz J. C., Casstevens T. M., Lu F., Harriman J., Elshire R. J., Sun Q. & Buckler E. S., 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS One*, 9 (2): e90346.
- Goddard M. E. & Hayes B. J., 2007. Genomic selection. *J Anim. Breed. Genet.*, 124 (6): 323-330.
- Goh K., 2000. Climatic requirements of the oil palm for high yields. *In: Managing oil palm for high yields: agronomic principles*. Goh K. (ed). Kuala Lumpur: 1-17.
- Goodnight C. J. & Kliman R. M., 2016. Gene Interactions in Evolution. *In: Kliman R. M. (ed). Encyclopedia of Evolutionary Biology*. Academic Press, Oxford: 104–109.
- Grattapaglia D., 2014. Breeding forest trees by genomic selection: current progress and the way forward. *In: Tuberosa, R., Graner, A., & Frison, E. (Eds). Genomics of plant genetic resources*. Springer, Dordrecht: 651–682.
- Grattapaglia D., Silva-Junior O. B., Resende R. T., Cappa E. P., Müller B. S., Tan B., Isik F., Ratcliffe B. & El-Kassaby Y. A., 2018. Quantitative genetics and genomics converge to accelerate forest tree breeding. *Front. Plant Sci.*, 9: 1693-1703.
- Habier D., Fernando R. & Dekkers J., 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177 (4): 2389–2397.
- Habier D., Fernando R. L., Kizilkaya K. & Garrick D. J., 2011. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12 (1): 186.
- Hardon J., Corley R. & Lee C., 1987. Breeding and selecting the oil palm. *In: Abbott A. J. & Atkin R. K. (eds). Improving vegetatively propagated crops*. Academic Press, London: 63-81.
- Hartley C. W. S., 1988. *The oil palm (Elaeis guineensis Jacq.)*. Longman Scientific & Technical, New York, 781 p.
- Hayes B. & Goddard M., 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.*, 33 (3): 209–229.
- He J., Zhao X., Laroche A., Lu Z.-X., Liu H. & Li Z., 2014. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.*, 5: 484-492.
- Heffner E. L., Sorrells M. E. & Jannink J.-L., 2009. Genomic selection for crop improvement. *Crop Sci.*, 49 (1): 1–12.
- Henry P., 1958. Croissance et développement chez *Elaeis guineensis* Jacq. de la germination a la première floraison. *Rev. Générale Bot.*, 66: 5–34.
- Hickey J. M., Kinghorn B. P., Tier B., van der Werf J. H. & Cleveland M. A., 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genet. Sel. Evol.*, 44 (1): 9.
- Hu X., 2015. A comprehensive comparison between ANOVA and BLUP to evaluate location-specific genotype effects for rape cultivar trials with random locations. *Field Crops Res.*, 179 144–149.
- Ibáñez-Escriche N., Fernando R., Toosi A. & Dekkers J., 2009. Genomic selection of purebreds for crossbred performance. *Genet. Sel. Evol.*, 41 (1): 12.

- Ithnin M., Xu Y., Marjuni M., Serdari N. M., Amiruddin M. D., Low E.-T. L., Tan Y.-C., Yap S.-J., Ooi L. C. L., Nookiah R., Singh R. & Xu S., 2017. Multiple locus genome-wide association studies for important economic traits of oil palm. *Tree Genet. Genomes*, 13 (5): 103-117.
- Jacob F., Cros D., Cochard B. & Durand-Gasselin T., 2017. Agrigenomics in the breeder's toolbox: latest advances towards an optimal implementation of genomic selection in oil palm. *In: International Seminar on 100 Years of Technological Advancement in Oil Palm Breeding & Seed Production*. Kuala Lumpur, 13 Nov 2017: 1-21.
- Jacquemard J. C., Baudoin L. & Noiret J. M., 1997. Le palmier à huile. *In: Charrier A., Jacquot M., Hamon S. & Nicolas D. (eds). L'amélioration des plantes tropicales*. CIRAD et ORSTOM, Paris: 507–531.
- Jacquemard J.-C., 1995. *Le palmier à huile*. Maisonneuve et Larose, Paris, 207 p.
- Jacquemard J.-C., 2012. *Le palmier à huile*. Quae, Versailles-Wageningen-Gembloux, 240 p.
- Jaligot E., Rival A., Beulé T., Dussert S. & Verdeil J.-L., 2000. Somaclonal variation in oil palm (*Elaeis guineensis* Jacq.): the DNA methylation hypothesis. *Plant Cell Rep.*, 19 (7): 684–690.
- Jourdan C. & Rey H., 1997. Architecture and development of the oil-palm (*Elaeis guineensis* Jacq.) root system. *Plant Soil*, 189 (1): 33–48.
- Junaidah J., Rafii M., Chin C. & Saleh G., 2011. Performance of Tenera oil palm population derived from crosses between deli Dura and Pisifera from different sources on inland soils. *J. Oil Palm Res.*, 23: 1210–1221.
- Kempthorne O., 1954. The correlation between relatives in a random mating population. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.*, 143 (910): 103–113.
- Kempthorne O., 1955. The theoretical values of correlations between relatives in random mating populations. *Genetics*, 40 (2): 153.
- Kilian A., Wenzl P., Huttner E., Carling J., Xia L., Blois H., Caig V., Heller-Uszynska K., Jaccoud D. & Hopper C., 2012. Diversity arrays technology: a generic genome profiling technology on open platforms. *In: Pompanon F. & Bonin A. (eds). Data production and analysis in population genomics*. Springer, Totowa: 67–89.
- Kwong Q. B., Ong A. L., Teh C. K., Chew F. T., Tammi M., Mayes S., Kulaveerasingam H., Yeoh S. H., Harikrishna J. A. & Appleton D. R., 2017a. Genomic selection in commercial perennial crops: applicability and improvement in oil palm (*Elaeis guineensis* Jacq.). *Scientific Reports*, 7 (1): 1–9.
- Kwong Q. B., Teh C. K., Ong A. L., Chew F. T., Mayes S., Kulaveerasingam H., Tammi M., Yeoh S. H., Appleton D. R. & Harikrishna J. A., 2017b. Evaluation of methods and marker Systems in Genomic Selection of oil palm (*Elaeis guineensis* Jacq.). *BMC Genet.*, 18 (107): 1–9.
- Kwong Q. B., Teh C. K., Ong A. L., Heng H. Y., Lee H. L., Mohamed M., Low J. Z.-B., Apparow S., Chew F. T., Mayes S., Kulaveerasingam H., Tammi M. & Appleton D. R., 2016. Development and validation of a high-density SNP genotyping array for African oil palm. *Mol. Plant*, 9 (8): 1132–1141.
- Lange K., 2003. *Mathematical and statistical methods for genetic analysis*. Springer Science & Business Media, New York, 346 p.

- Langmead B. & Salzberg S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9 (4): 357.
- Leroy T., Montagnon C., Cilas C., Yapo A., Charmetant P. & Eskes A., 1997. Reciprocal recurrent selection applied to *Coffea canephora* Pierre. III. Genetic gains and results of first cycle intergroup crosses. *Euphytica*, 95 (3): 347–354.
- Li C., Bai G., Chao S. & Wang Z., 2015. A high-density SNP and SSR consensus map reveals segregation distortion regions in wheat. *BioMed Res. Int.*, 2015.
- Lopes M. S., Bovenhuis H., Hidalgo A. M., Van Arendonk J. A., Knol E. F. & Bastiaansen J. W., 2017. Genomic selection for crossbred performance accounting for breed-specific effects. *Genet. Sel. Evol.*, 49 (1): 51.
- Lorenz A. J., Chao S., Asoro F. G., Heffner E. L., Hayashi T., Iwata H., Smith K. P., Sorrells M. E. & Jannink J.-L. Donald L. Sparks, 2011. Genomic Selection in Plant Breeding: Knowledge and Prospects. In: Donald L. S. (ed.) *Advances in Agronomy*. Academic Press, Amsterdam-Boston-Heidelberg-London-New York-Oxford-Paris-San Diego-San Francisco-Singapore-Sydney-Tokyo: 77–123.
- Luyindula N., Mantantu N., Dumortier F. & Corley R. H. V., 2005. Effects of inbreeding on growth and yield of oil palm. *Euphytica*, 143 (1–2): 9–17.
- Lynch M. & Walsh B., 1998. *Genetics and analysis of quantitative traits*. Sinauer Sunderland, MA, Sunderland 980 p.
- Lyttle T. W., 1991. Segregation distorters. *Annu. Rev. Genet.*, 25 (1): 511–581.
- Lyttle T. W., 1993. Cheaters sometimes prosper: distortion of mendelian segregation by meiotic drive. *Trends Genet.*, 9 (6): 205–210.
- Malécot G., 1948. *Les mathématiques de l'hérédité*. Masson et Cie, Paris, 60 p.
- Marchal A., Legarra A., Tisné S., Carasco-Lacombe C., Manez A., Suryana E., Omoré A., Durand-Gasselín T., Sánchez L., Bouvet J.-M. & Cros D., 2016. Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests. *Mol. Breed.*, 36 (2): 1–13.
- Masani M. Y. A., Izawati A. M. D., Rasid O. A. & Parveez G. K. A., 2018. Biotechnology of oil palm: Current status of oil palm genetic transformation. *Biocatal. Agric. Biotechnol.*, 15: 335–347.
- Mayes S., 2020. The History and economic importance of the oil palm. In: Ithnin M. & Kushairi A. (eds). *The oil palm genome*. Springer, Cham: 1–8.
- Mendiburu F., 2016. *agricolae: Statistical Procedures for Agricultural Research*. R package.
- Meunier J., 1969. Etude des populations naturelles d'*Elaeis guineensis* en Côte-d'Ivoire. *Oléagineux*, 24(4): 195-201.
- Meunier J. & Boutin D., 1975. L'*Elaeis melanococca* et l'hybride *Elaeis melanococca* x *Elaeis guineensis*. *Oléagineux*, 30 (1): 5-8.
- Meunier J. & Gascon J., 1972. Le schéma général d'amélioration du palmier à huile à l'IRHO. *Oléagineux*, 27 (1): 1–12.
- Meuwissen T. H. E., Hayes B. J. & Goddard M. E., 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157 (4): 1819–1829.

- Misztal I., Aguilar I., Johnson D., Legarra A., Tsuruta S. & Lawlor T., 2009. A unified approach to utilize phenotypic, full pedigree and genomic information for a genetic evaluation of Holstein final score. *Interbull Bull.* 40: 240.
- Mrode R. A., 2005. *Linear models for the prediction of animal breeding values*. CABI, Oxfordshire, UK, 344 p.
- Muranty H., Jorge V., Bastien C., Lepoittevin C., Bouffier L. & Sanchez L., 2014. Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. *Tree Genet. Genomes*, 10 (6): 1491–1510.
- Neale M., Boker S., Xie G. & Maes H., 2003. *Mx: Statistical modeling*. Richmond, VA: Department of Psychiatry. *Va. Inst. Psychiatr. Behav. Genet. Va. Commonw. Univ.*
- Noh A., Rafii M., Saleh G., Kushairi A. & Latif M., 2012. Genetic performance and general combining ability of oil palm Deli dura x AVROS pisifera tested on inland soils. *Sci. World J.*, 2012: 1-8.
- Nouy B., Jacquemard J.-C., Suryana E., Potier F., Konan K. E. & Durand-Gasselin T., 2006. The expected and observed characteristics of several oil palm (*Elaeis guineensis* Jacq.) clones. *In: International Oil Palm Conference*. Bali, Indonesia, 19-23 July: 1-17.
- Nyouma A., Bell J. M., Jacob F. & Cros D., 2019. From mass selection to genomic selection: one century of breeding for quantitative yield components of oil palm (*Elaeis guineensis* Jacq.). *Tree Genet. Genomes*, 15 (5): 69.
- Nyouma A., Bell J. M., Jacob F., Riou V., Manez A., Pomiès V., Nodichao L., Syahputra I., Affandi D. & Cochard B., 2020. Genomic predictions improve clonal selection in oil palm (*Elaeis guineensis* Jacq.) hybrids. *Plant Sci.*, 299: 1-12.
- Okoye M., Okwuagwu C. & Uguru M., 2009. Population improvement for fresh fruit bunch yield and yield components in oil palm (*Elaeis guineensis* Jacq.). *Am.-Eurasian J. Sci. Res.*, 4 (2): 59–63.
- Okwuagwu C., Okoye M. N., Okolo E., Ataga C. & Uguru M., 2008. Genetic variability of fresh fruit bunch yield in Deli/dura x tenera breeding populations of oil palm (*Elaeis guineensis* Jacq.) in Nigeria. *J. Trop. Agric.*, 46: 52–57.
- Ong-Abdullah M., Ordway J. M., Jiang N., Ooi S., Kok S.-Y., Sarpan N., Azimi N., Hashim A. T., Ishak Z. & Rosli S. K., 2015. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature*, 525 (7570): 533-537.
- Oraguzie N. C., Rikkerink E. H. A., Gardiner S. E. & de Silva H. N., 2007. *Association Mapping in Plants*. Springer, New York, 277 p.
- Pérez P., Crossa J., De Los Campos G. & Gianola D., 2010. Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. *Plant Genome*, 3 (2): 106–116.
- Periasamy A., Gopal K. & Soh A., 2002. Productivity improvements in seed processing techniques for commercial oil palm seed production. *Planter*, 78 (917): 429–442.
- Piepho H., Möhring J., Melchinger A. & Büchse A., 2008. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161 (1–2): 209–228.
- Pootakham W., Jomchai N., Ruang-areerate P., Shearman J. R., Sonthirod C., Sangsrakru D., Tragoonrun S. & Tangphatsornruang S., 2015. Genome-wide SNP discovery and identification

- of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics*, 105 (5): 288–295.
- Potier F., Nouy B., Flori A., Jacquarmard J., Edyana Suryana H. & Durand-Gasselín T., 2006. Yield potential of oil palm (*Elaeis guineensis* Jacq.) clones: preliminary results observed in the Aek Loba genetic block in Indonesia. *In: International Seminar on Yield Potential in Oil Palm: Yield potential in oil palm II*. Phuket, 27-28 Nov 2006: 1-20.
- Powell J. E., Visscher P. M. & Goddard M. E., 2010. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat. Rev. Genet.*, 11: 800–805.
- Pszczola M., Strabel T., Mulder H. & Calus M., 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.*, 95 (1): 389–400.
- Purba A. R., Flori A., Baudouin L. & Hamon S., 2001. Prediction of oil palm (*Elaeis guineensis* Jacq.) agronomic performances using the best linear unbiased predictor (BLUP). *Theor. Appl. Genet.*, 102 (5): 787–792.
- Purseglove J., 1976. The origins and migrations of crops in tropical Africa. *In: Harlan J. R., De Wet J. M. J., Stemler A. B. L. (eds). Origins of African plant domestication*. Mouton Publishers, The Hague: 291–310.
- R Core Team, 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rafflegeau S., 2008. *Dynamique d'implantation et conduite technique des plantations villageoises de palmier à huile au Cameroun: facteurs limitants et raisons des pratiques*. PhD thesis, AgroParisTech, 148 p.
- Rafii M. Y., Isa Z. A., Kushairi A., Saleh G. B. & Latif M. A., 2013. Variation in yield components and vegetative traits in Malaysian oil palm (*Elaeis guineensis* Jacq.) dura×pisifera hybrids under various planting densities. *Ind. Crops Prod.*, 46: 147–157.
- Rajanaidu N., 1986. The oil palm (*Elaeis guineensis*) collections in Africa. *In: International Workshop on Oil Palm Germplasm and Utilisation*. Bangi-Selangor, 26-27 March 1985: 59-83.
- Rajanaidu N., Tan Y. P., Ong E. C. & Lee C. H., 1986. The performance of inter-origin commercial D×P planting material. *Perform. Inter-Orig. Commer. D×P Plant. Mater.*, (10): 155–161.
- Rao V. & Kushairi A., 1999. Quality of oil palm planting material. *In: Proceeding of the 1996 Seminar on Sourcing of Oil Palm Planting Materials for Local and Overseas Joint Ventures*. Rajanaidu N. & Jalani B.S. (eds). Bangi, 188–197.
- Reif J. C., Gumpert F.-M., Fischer S. & Melchinger A. E., 2007. Impact of interpopulation divergence on additive and dominance variance in hybrid populations. *Genetics*, 176 (3): 1931–1934.
- Revelle W., 2018. *psych: Procedures for psychological, psychometric, and personality research, R package 1.8. 4*. Northwestern University, Evanston, Illinois.
- Rincent R., Charcosset A. & Moreau L., 2017. Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theor. Appl. Genet.*, 130 (11): 1–17.
- Rincent R., Laloë D., Nicolas S., Altmann T., Brunel D., Revilla P., Rodriguez V. M., Moreno-Gonzalez J., Melchinger A. & Bauer E., 2012. Maximizing the reliability of genomic selection by

- optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics*, 192 (2): 715–728.
- Rival A. & Levang P., 2014. *Palms of controversies: oil palm and development challenges*. CIFOR, Jakarta, 58 p.
- Rosenquist E., 1986. The genetic base of oil palm breeding populations. *In*: International Workshop on Oil Palm Germplasm and Utilization. Bangi-Selangor, March 26-27: 27-56.
- Sansaloni C., Petroli C., Jaccoud D., Carling J., Detering F., Grattapaglia D. & Kilian A., 2011. Diversity Arrays Technology (DART) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. *BMC Proc.*, 5 (7): 1-2.
- Schnell F. & Cockerham C., 1992. Multiplicative vs. arbitrary gene action in heterosis. *Genetics*, 131 (2): 461–469.
- Severson A. L., Carmi S. & Rosenberg N. A., 2019. The effect of consanguinity on between-individual identity-by-descent sharing. *Genetics*, 212 (1): 305–316.
- Singh R., Low E.-T. L., Ooi L. C.-L., Ong-Abdullah M., Chin T. N., Nagappan J., Nookiah R., Amiruddin M. D., Rosli R. & Manaf M. A. A., 2013a. The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK. *Nature*, 500 (7462): 340.
- Singh R., Low E.-T. L., Ooi L. C.-L., Ong-Abdullah M., Nookiah R., Ting N.-C., Marjuni M., Chan P.-L., Ithnin M. & Manaf M. A. A., 2014. The oil palm VIRESCENS gene controls fruit colour and encodes a R2R3-MYB. *Nat. Commun.*, 5 (1): 1–8.
- Singh R., Ong-Abdullah M., Low E.-T. L., Manaf M. A. A., Rosli R., Nookiah R., Ooi L. C.-L., Ooi S., Chan K.-L. & Halim M. A., 2013b. Oil palm genome sequence reveals divergence of interfertile species in old and new worlds. *Nature*, 500 (7462): 335-339.
- Soh A., 1994. Ranking parents by best linear unbiased prediction (BLUP) breeding values in oil palm. *Euphytica*, 76 (1): 13–21.
- Soh A., 1999. Breeding plans and selection methods in oil palm. *In*: Symposium on the science of oil palm breeding. Rajanaidu N. & Jalani B.S. (eds). Montpellier: 65-95 .
- Soh A. C., Mayes S. & Roberts J. A., 2017. *Oil Palm Breeding: Genetics and Genomics*. CRC Press, Boca Raton, 446 p.
- Soh A. C., Wong C. K., Ho Y. W. & Choong C. W. Vollmann J. & Rajcan I., 2010. Oil Palm. *In*: Vollmann J. & Rajcan I. (eds). *Oil Crops*. Springer, New York: 333–367.
- Soh A., Gan H., Wong G., Hor T. & Tan C., 2003. Estimates of within family genetic variability for clonal selection in oil palm. *Euphytica*, 133 (2): 147–163.
- Soh A., Wong G., Hor T., Tan C. & Chew P., 2003. Oil palm genetic improvement. *Plant Breed. Rev.*, 22: 165–220.
- Stearns S., 1992. *The Evolution of Life Histories*. Oxford: Oxford Univ. Press. 249 p.
- Steiger J. H., 1980. Tests for comparing elements of a correlation matrix. *Psychol. Bull.*, 87 (2): 245–251.

- Stock J., Bennewitz J., Hinrichs D. & Wellmann R., 2020. A review of genomic models for the analysis of livestock crossbred data. *Front. Genet.*, 11: 1-10.
- Stuber C. & Cockerham C. C., 1966. Gene effects and variances in hybrid populations. *Genetics*, 54 (6): 1279.
- Su G., Christensen O. F., Ostersen T., Henryon M. & Lund M. S., 2012. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. *PloS One*, 7 (9): e45293.
- Syed R., 1982. Insect pollination of oil palm: feasibility of introducing *Elaeidobius* spp. into Malaysia. *In: Proceedings of the International Conference on Oil Palm in Agriculture in the Eighties: The oil palm in the eighties*. Kuala Lumpur, 1: 263–289.
- Taylor D. R. & Ingvarsson P. K., 2003. Common features of segregation distortion in plants and animals. *Genetica*, 117 (1): 27–35.
- Technow F., Riedelsheimer C., Schrag Tobias A. & Melchinger Albrecht E., 2012. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor. Appl. Genet.*, 125 (6): 1181–1194.
- Technow F., Schrag T. A., Schipprack W., Bauer E., Simianer H. & Melchinger A. E., 2014. Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics*, 197 (4): 1343-1355.
- Ting N.-C., Jansen J., Mayes S., Massawe F., Sambanthamurthi R., Cheng-Li O., Chin C., Arulandoo X., Seng T.-Y., Alwee S., Ithinin M. & Singh R., 2014. High density SNP and SSR-based genetic maps of two independent oil palm hybrids. *BMC Genomics*, 15 (1): 309.
- Ting N.-C., Mayes S., Massawe F., Sambanthamurthi R., Jansen J., Syed Alwee S. S. R., Seng T.-Y., Ithinin M. & Singh R., 2018. Putative regulatory candidate genes for QTL linked to fruit traits in oil palm (*Elaeis guineensis* Jacq.). *Euphytica*, 214 (11): 214.
- Tisné S., Denis M., Cros D., Pomiès V., Riou V., Syahputra I., Omoré A., Durand-Gasselín T., Bouvet J.-M. & Cochard B., 2015. Mixed model approach for IBD-based QTL mapping in a complex oil palm pedigree. *BMC Genomics*, 16 (1): 1–12.
- Toro M., Fernández J., Shaat I. & Mäki-Tanila A., 2011. Assessing the genetic diversity in small farm animal populations. *Animal*, 5 (11): 1669–1683.
- Trustring G. & Williamson J., 1961. The correlations between relatives in a random mating diploid population. *In: Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge University Press, Cambridge, 57 (2): 315–320.
- Ukrainetz N. K. & Mansfield S. D., 2020. Prediction accuracy of single-step BLUP for growth and wood quality traits in the lodgepole pine breeding program in British Columbia. *Tree Genet. Genomes*, 16 (5): 1–13.
- VanRaden P. M., 2007. Genomic measures of relationship and inbreeding. *Interbull Bull.*, 37: 33–36.
- VanRaden P. M., 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91 (11): 4414–4423.
- Varshney R. K., Roorkiwal M. & Sorrells M. E., 2017. *Genomic selection for crop improvement*. Springer International Publishing, Cham, 258 p.

- Verrier E., Brabant P. & Gallais A., 2001. *Faits et concepts de base en génétique quantitative*. INA Paris-Grignon, Paris, 132 p.
- Visscher P. M., Medland S. E., Ferreira M. A. R., Morley K. I., Zhu G., Cornes B. K., Montgomery G. W. & Martin N. G., 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.*, 2 41.
- Vitezica Z. G., Varona L., Elsen J.-M., Miształ I., Herring W. & Legarra A., 2016. Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. *Genet. Sel. Evol.*, 48 (1): 6.
- White T. L. & Hodge G. R., 1989. *Predicting breeding values with applications in forest tree improvement*. Springer, Netherlands, 367 p.
- Wei M., Van der Werf J. H. J. & Brascamp E. W., 1991. Relationship between purebred and crossbred parameters. *J. Anim. Breed. Genet.*, 108 (1–6): 262–269.
- White T. L. & Hodge G. R., 1989. *Predicting breeding values with applications in forest tree improvement*. Springer, Netherlands, 367 p.
- Wiggans G. R., Cole J. B., Hubbard S. M. & Sonstegard T. S., 2017. Genomic Selection in Dairy Cattle: The USDA Experience. *Annu. Rev. Anim. Biosci.*, 5 (1): 309–327.
- Wong C. & Bernardo R., 2008. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor. Appl. Genet.*, 116 (6): 815–824.
- Wray N. & Visscher P., 2008. Estimating trait heritability. *Nat. Educ.*, 1 (1): 29.
- Wright S., 1921. Systems of mating. I. The biometric relations between parent and offspring. *Genetics*, 6 (2): 111-123.
- Wright S., 1922. Coefficients of inbreeding and relationship. *Amer. Nat.*, 56: 330–338.
- Xavier A., Muir W. M., Craig B. & Rainey K. M., 2016. Walking through the statistical black boxes of plant breeding. *Theor. Appl. Genet.*, 129 (10): 1933–1949.
- Xiang T., Nielsen B., Su G., Legarra A. & Christensen O. F., 2016. Application of single-step genomic evaluation for crossbred performance in pig. *J. Anim. Sci.*, 94 (3): 936–948.
- Xian-Liang S., Xue-Zhen S. & Tian-Zhen Z., 2006. Segregation distortion and its effect on genetic mapping in plants. *Chin. J. Agric. Biotechnol.*, 3 (3): 163–170.
- Xu S., 2013. *Principles of statistical genomics*. Springer, New York-Heidelberg-Dordrecht-London, 428 p.
- Xu X., Li L., Dong X., Jin W., Melchinger A. E. & Chen S., 2013. Gametophytic and zygotic selection leads to segregation distortion through in vivo induction of a maternal haploid in maize. *J. Exp. Bot.*, 64 (4): 1083–1096.
- Zeng J., Toosi A., Fernando R. L., Dekkers J. C. & Garrick D. J., 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genet. Sel. Evol.*, 45 (1): 1–17.
- Zeven A., 1964. On the origin of the oil palm (*Elaeis guineensis* Jacq.). *Grana*, 5 (1): 121–123.
- Zuo J.-F., Niu Y., Cheng P., Feng J.-Y., Han S.-F., Zhang Y.-H., Shu G., Wang Y. & Zhang Y.-M., 2019. Effect of marker segregation distortion on high density linkage map construction and QTL mapping in Soybean (*Glycine max* L.). *Heredity*, 123 (5): 579–592.

APPENDICES

Appendix 1. Objectives and corresponding published papers.

Objectives		Published papers
general	specifics	
to evaluate empirically the interest of using genomic data from A × B hybrid individuals for the genomic approach applied to oil palm	to evaluate the efficiency of genomic selection for clonal selection, using ortets of known clonal value to validate genomic predictions	Genomic predictions improve clonal selection in oil palm (<i>Elaeis guineensis</i> Jacq.) hybrids
	to investigate the effect of the genotyping strategy to optimize prediction accuracy	Improving accuracy of genomic predictions in an outcrossing species with hybrid cultivars between heterozygote parents: case study of oil palm (<i>Elaeis guineensis</i> Jacq.) (accepted)
		From mass selection to genomic selection: one century of breeding for quantitative yield components of oil palm (<i>Elaeis guineensis</i> Jacq.)

Appendix 2. Logical framework of objective 1: evaluation of the efficiency of genomic selection for clonal selection, using ortets of known clonal value to validate genomic predictions

Materials	Methods	Results	Conclusion
<p>Training set: 300 Deli × La Mé crosses phenotyped for eight yield components (average individuals per cross is 67 individuals for bunch production and 44 for bunch quality).</p> <p>Validation set: 42 Deli × La Mé ortets (average of 69 ramets per ortets for production traits and 34 ramets for quality traits)</p> <p>Number of SNP: 15,054</p>	<p>Individuals of the training set phenotyped for AFW, FB, PF, OP, NF, BN, ABW and FFB.</p> <p>Molecular data were obtained by GBS</p> <p>Imputation of missing SNP data and phasing were carried out with Beagle 4.0.</p> <p>To quantify how the characteristics of the SNP dataset (maximum percentage of missing data allowed per SNP, and resulting number of SNPs) affected the GS accuracy, genomic predictions were computed using different SNP datasets.</p> <p>Two approaches of marker modeling were considered: one taking into account the parental origin of marker alleles, PSAM, or not, ASGM</p> <p>- ASGM: $y = X\beta + Z_1g_i + Z_2g_{Deli \times LM} + Z_3b + Z_4p + \varepsilon$</p> <p>- PSAM: $y = X\beta + Z_1g_{Deli} + Z_2g_{LM} + Z_3g_{Deli \times LM} + Z_4b + Z_5p + \varepsilon$</p> <p>Calculation of genetic values</p> <p>$\hat{g}_{Deli} + \hat{g}_{LM}$</p> <p>Prediction accuracy of GS</p> <p>$r_{GS} = Cor(\hat{g}_{true}, \hat{g}_{SG})$</p> <p>Pairwise comparisons of prediction accuracies among models were made for each trait using the Hotelling–Williams <i>t</i>-test</p> <p>The differences in accuracy between ASGM and PSAM were explained using the distribution of the MAF and of the frequency of the alternate allele in Deli and La Mé, as well as the correlation among populations for each of these two parameters.</p> <p>Determination of reference clonal value predicted by the models</p> <p>to validate the different prediction models, clonal genetic values were obtained for each clone from the phenotypic data collected on their ramets.</p>	<p>MAF ranged from 0 to 0.5 for both La Mé and Deli populations and the average was 0.1 for La Mé and 0.07 for Deli. 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. 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 %).</p> <p>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.</p> <p>Prediction accuracies were ranging from 0.08 to 0.70 for ortet candidates without data records, depending on trait, SNP dataset and modeling</p> <p>ASGM was better (more robust over traits and SNP datasets, and simpler), although PSAM could noticeably improve prediction accuracies for some traits. The number of SNPs had to reach 7,000, while the percentage of missing data per SNP was of secondary importance for modeling approaches.</p> <p>GS prediction accuracies were higher than those of PS for most of the traits.</p>	<p>This study 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.</p>

Appendix 3. Logical framework of objective 2: investigation of the effect of the genotyping strategy to optimize prediction accuracy.

Materials	Methods	Results	Conclusion
<p>Training set: 350 hybrid crosses phenotyped for nine yield components (average number of individuals per cross of 64 for bunch production and 44 for bunch quality) + 400 training hybrid individuals.</p> <p>Validation set: 213 hybrid crosses (average number of individuals per cross of 63 for production traits and 48 for quality traits)</p> <p>Number of SNP: 21,458</p>	<p>Individuals of the training set phenotyped for AFW, FB, PF, OP, NF, OER, BN, ABW and FFB.</p> <p>Molecular data were obtained by GBS</p> <p>Imputation of missing SNP data and phasing were carried out with Beagle 4.0.</p> <p>effects of the SNP dataset i.e., density and percentage of missing data</p> <p>Two approaches of marker modeling were considered: one taking into account the parental origin of marker alleles, PSAM, or not, ASGM</p> <p>- ASGM: $y = X\beta + Z_g g_g + Z_b b + Z_p p + \varepsilon$</p> <p>- PSAM: $y = X\beta + Z_A g_A + Z_B g_B + Z_b b + Z_p p + \varepsilon$</p> <p>Calculation of genetic values $\hat{g}_A + \hat{g}_B$</p> <p>Determination of reference value of validation hybrid crosses To validate the different prediction models, the true genetic value of the validation hybrid crosses, termed reference genetic value, was computed from the phenotypic data of their hybrid individuals</p> <p>Prediction accuracy of GS $r_{GS} = Cor(\hat{g}_{true}, \hat{g}_{GS})$</p> <p>The comparison of models was carried out an ANOVA using <i>agricolae</i> R package</p>	<p>Prediction accuracies ranged from 0.15–0.89 depending on trait, model and genotyping strategy.</p> <p>GS prediction accuracies increased on average by 5% when training was done with genomic data of hybrid individuals and parents compared with only parental genomic data.</p> <p>On average over traits, G_ASGM_Par+Hyb with a prediction accuracy of 0.53, was significantly higher than G_ASGM_Par with 0.50</p> <p>GS prediction accuracies increased on average by 3% with ASGM compared to PSAM.</p>	<p>Adding genomic data of hybrid individuals when training the model increased GS accuracy</p> <p>ASGM was the best model (giving the highest prediction accuracies on average over traits)</p> <p>G_ASGM_Par+Hyb with a prediction accuracy of 0.53 was the best GS approach</p> <p>ASGM approach is recommended for oil palm data, as it gives higher prediction accuracies on average over traits, performs best on more traits and is more robust over populations and SNP datasets than the PSAM approach.</p>

Appendix 4. Generation of SNP molecular data (Cros *et al.*, 2017).

DNA extraction was performed by ADNid (www.adnid.fr) on lyophilized tissue from the youngest opened leaf of each individual, using a modified mixed alkyltrimethylammonium bromide (MATAB) protocol. GBS was conducted on the DNA extracts by a company called DArT (www.diversityarrays.com) using their DArTseq™ protocol (Kilian *et al.*, 2012), which combined complexity reduction of the genome and next generation sequencing (Baird *et al.*, 2008; Pootakham *et al.*, 2015). DNA samples were processed in digestion/ligation reactions mainly as per Kilian *et al.* (2012) but using two adaptors corresponding to the *Pst*I and *Hha*I restriction enzyme overhangs and moving the assay on the sequencing platform as described by Sansaloni *et al.* (2011). The *Pst*I-compatible adapter was designed to include the Illumina flow cell attachment sequence, the sequencing primer sequence and the “staggered”, varying length barcode region, similar to the sequence reported by Elshire *et al.* (2011). The reverse adapter contained the flowcell attachment region and the *Hha*I-compatible overhang sequence. Only *Pst*I-*Hha*I mixed fragments were effectively amplified in 30 rounds of PCR using the following reaction conditions: (1) 94 °C for 1 min, (2) 30 cycles at 94 °C for 20 s, 58 °C for 30 s, 72 °C for 45 s and (3) 72 °C for 7 min. Next, PCR equimolar amounts of amplification products from each sample in the 96-well microtiter plate were bulked and applied to c-Bot (Illumina) bridge PCR followed by sequencing on Illumina HiSeq2500. Single read sequencing was run for 77 cycles.

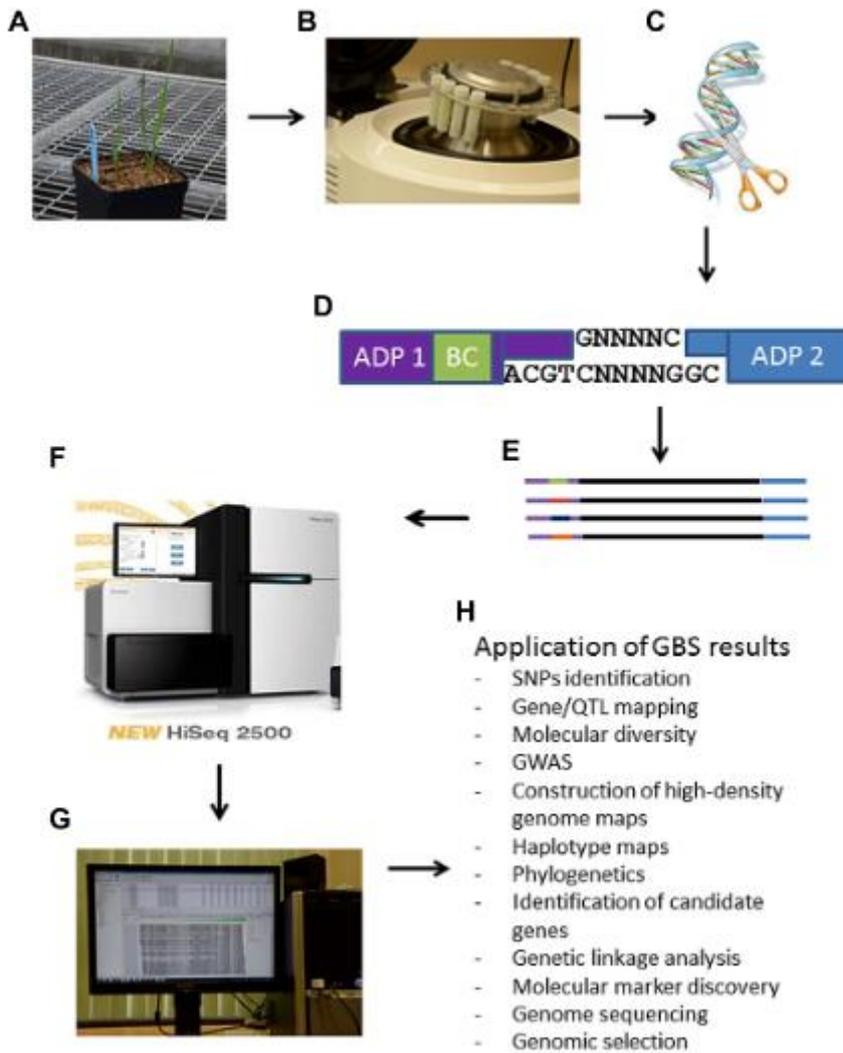
The GBS analysis pipeline implemented in Tassel GBS version 5.2.29 (Glaubitz *et al.*, 2014) was used to call SNPs according to the parameters listed in Table S1. From the total number of good barcoded reads (152,020,019 out of 238,493,056), the pipeline found 476,589 tags, aligned with Bowtie2 software. The tag mapping and the polymorphism calling identified 109,201 polymorphic sites. The data were further processed with VCFtools (Danecek *et al.*, 2011). Indels and SNPs that were not biallelic were discarded. Data points with a sequencing depth of less than five were set to missing. SNPs with more than 50% missing data were discarded. Using a custom R script (R Core Team, 2017), the SNPs appearing as outliers in terms of mean depth (i.e. higher than 500) were discarded, as it was assumed this could indicate duplication in the genome. This resulted in 19,432 SNPs. The molecular dataset was split into two, one for Group A and the other for Group B. The SNPs that mapped on the unassembled part of the genome were discarded, as the imputation of sporadic missing data required known positions. Mendelian segregation between parents and offspring was checked and the

inconsistent data points were set to missing. The SNP homozygotes or with more than 5% of Mendelian inconsistencies in a parental group were discarded from this group.

Table S1. Tassel v5.2.29 GBS pipeline used to process raw sequence data.

Step_plugin	Parameters	Results	Value	%
00 (raw fastq data)		Number of reads in lanes	238,493,056	
01_GBSSeqToTagDB	ePstI c20 kmerL68 minKmerL20 mnQS20	Number of correct barcoded reads	152,020,019	63.7
01_GBSSeqToTagDB	ePstI c20 kmerL68 minKmerL20 mnQS20	Number of tags	476,589	
02_TagExportToFastq	c1	Export tags to fastq	476,589	
03_BowtieToSAM	very-sensitive-local	Number of tags aligned once	243,794	51.2
03_BowtieToSAM	very-sensitive-local	Number of tags aligned >1 time	77,160	16.2
04_SAMToGBSdb	aProp0 aLen0	Number of mapped tags	320,954	67.3
05_DiscoverySNP Caller	maxTagsCutSite68 mnLCov0.1 mnMAF0.0025 eR 0.01	Number of polymorphic sites	109,201	
05_DiscoverySNP Caller	maxTagsCutSite68 mnLCov0.1 mnMAF0.0025 eR 0.01	Number of alleles	230,100	
06_SNPQuality Profiler		Number of polymorphic sites	109,201	
07_ProdSNP Caller	ePstI kmerL68 mnQS0	Number of polymorphic sites	109,201	

Appendix 5. Steps of genotyping-by-sequencing (GBS) in plants (He et al., 2014).



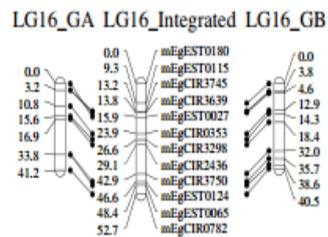
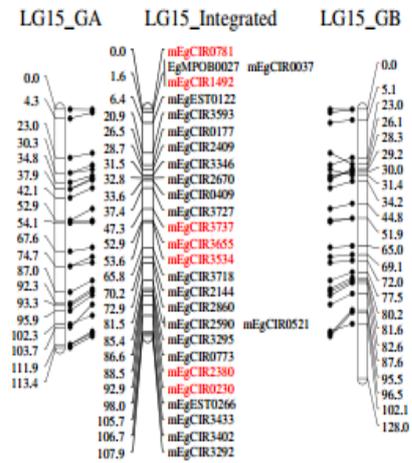
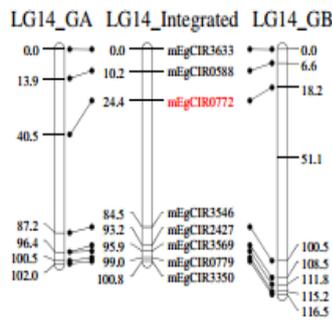
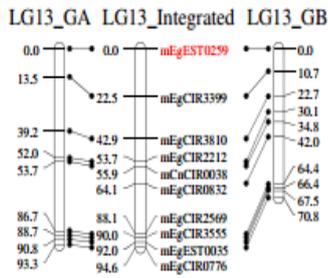
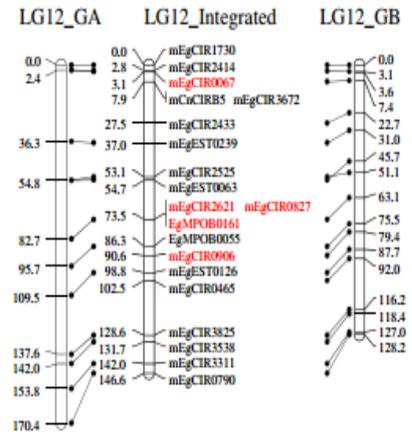
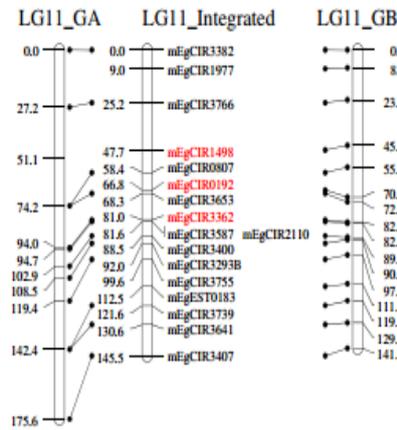
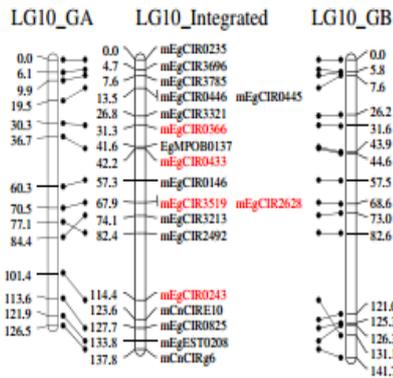
A: tissue is obtained from any plant species; B: ground leaf tissues for DNA isolation, quantification and normalization. At this step it is important to prevent any cross-contamination among samples; C: DNA digestion with restriction enzymes; D: ligations of adaptors (ADP) including a bar coding (BC) region in adapter 1 in random *Pst*I restricted DNA fragments; E: representation of different amplified DNA fragments with different bar codes from different biological samples. These fragments represent the GBS library; F: analysis of sequences from library on a NGS sequencer; G: bioinformatic analysis of NGS sequencing data; H: possible application of GBS results.

Appendix 6. Genetic map of oil palm.

Group A (GA, left), integrated (middle), and group B (GB, right). Marker names are indicated only at the right of the integrated genetic map, and map distances in centimorgans (cM) at the left of each genetic map. Lines between linkage groups (LG) show common markers between maps. Markers in black were positioned on the genome sequence while those in red could not be positioned.



Appendix 6 (continued)



Appendix 7. Published papers.



From mass selection to genomic selection: one century of breeding for quantitative yield components of oil palm (*Elaeis guineensis* Jacq.)

Achille Nyouma^{1,2} · Joseph Martin Bell¹ · Florence Jacob³ · David Cros^{2,4,5}

Received: 20 January 2019 / Revised: 27 May 2019 / Accepted: 15 July 2019 / Published online: 15 August 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

More efficient methods are required to breed oil palm (*Elaeis guineensis* Jacq.) for yield maximization in order to meet the increased demand for palm oil while limiting environmental impacts. This review article analyzes the evolution of breeding schemes for oil palm yield and its quantitative components and the changes expected to take place with genomic selection (GS). Genetic improvement of oil palm yield started in the 1920s through mass selection. Later, several disruptive improvements dramatically increased the rate of genetic progress: (1) understanding the heredity of fruit form and the adoption of *tenera*, with thicker mesocarp, in plantations; (2) the discovery of hybrid vigor and the adoption of modified reciprocal recurrent selection; and (3) clonal selection, exploiting intra-hybrid variability. In addition, the use of linear mixed models to estimate genetic values has made selection more efficient. Today, GS appears to be a new disruptive improvement that can speed up breeding schemes by avoiding field trials in some cycles and increase selection intensity by evaluating more candidates. The genetic potential for oil palm yield has increased considerably over one century of breeding. GS is expected to bring the rate of genetic progress to a previously unprecedented level. The future studies on oil palm GS will aim at making it efficient for all yield components. For this purpose, they should focus in particular on the optimization of training populations and on the improvement of prediction models. Minimizing environmental impacts will also require improvement in other aspects (resistance to diseases, cultural practices, etc.).

Keywords *Elaeis guineensis* · Hybrids · Reciprocal recurrent selection · Genomic selection · BLUP · Linear mixed model

Introduction

Oil palm (*Elaeis guineensis* Jacq.) is the most productive oil crop in the world, with annual production of more than 65 million tons of palm oil (USDA 2018). The world population is expected to be over nine billion by 2050, and the demand for palm oil to be between 120 and 156 million tons (Corley 2009; Rival and Levang 2014). Genetic improvement has a major role to play to meet this demand while minimizing environmental impacts. Indeed, a significant proportion of the increase in yield already achieved was due to breeding. In Malaysia, 70% of the increase in yield is attributed to genetic improvement, versus only 30% to improvement in cultural practices (Davidson 1993). The genetic progress in palm oil yield is currently estimated at around 1% and 1.5% per year, comparable to that of maize (Rival and Levang 2014, p. 39).

Oil palm originated from the Gulf of Guinea. It is a tree-like diploid ($2n = 2x = 32$ chromosomes) monocotyledon from the

Communicated by S. C. González-Martínez

✉ David Cros
david.cros@cirad.fr

¹ Department of Plant Biology, Faculty of Science, University of Yaoundé I, Yaoundé, Cameroon

² CETIC (African Center of Excellence in Information and Communication Technologies), University of Yaoundé I, Yaoundé, Cameroon

³ PalmElit SAS, 34980 Montferrier sur Lez, France

⁴ CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP, 34398 Montpellier, France

⁵ AGAP, CIRAD, INRA, Montpellier SupAgro, University of Montpellier, Montpellier, France

Arecaceae family (Jacquemard et al. 1997). There are three types based on fruit morphology: *dura* (*D*), whose fruits contain a thick-shelled nut; *pisifera* (*P*), which has no shell and is generally female sterile; and *tenera* (*T*) which has a thin shell and is female fertile. Oil palm is allogamous and artificial pollination allows controlled crosses (Fig. 1a). It has no natural means of vegetative propagation but cloning is possible in vitro by tissue culture in the laboratory. Palm oil yield (OY) in *E. guineensis* is a complex trait. The leaves are emitted successively and each bears an inflorescence bud in its axil, which, unless abortion occurs, produces an inflorescence, with alternating male and female cycles throughout the life of the plant (Demol 2002). The pollinated female inflorescence develops into a bunch. Fresh fruit bunch (FFB) production, which is one of the two main components of OY, results from the number of bunches (BN) and average bunch weight (BW). The second main component is the percentage of oil in the bunches (O/B), which can be broken down into more simple traits (Fig. 1b, c), i.e., the percentage of fruits in the bunch (F/B), the percentage of pulp or mesocarp per fruit (M/F), and the percentage of oil in the mesocarp (O/M).

Genetic improvement of oil palm yield started with mass selection in the 1920s (Demol 2002; Corley and Tinker 2016). An understanding of the genetic determinism of the fruit form was acquired in the 1930s (Beirnaert and Vanderweyen 1941). In the 1950s and 1960s, mass selection was replaced by more efficient breeding schemes, mainly of modified reciprocal recurrent selection type (MRRS), leading to interpopulation hybrid *T* cultivars (Corley and Tinker 2016; Soh et al. 2017).

Since the late 1970s, clonal varieties from tissue culture have also been produced (Corley and Tinker 2016, p. 208; Soh et al. 2017, p. 193). The current period is marked by two new changes. The first is the adoption of more efficient statistical methods to estimate the genetic value of the selection candidates, with a shift from analysis of variance (ANOVA) to the BLUP (best linear unbiased predictor) method (Henderson 1950, 1984). Although this shift started several decades ago, current literature indicates that it is still underway. The second very recent change is the use of approaches that take advantage of genomic data. For quantitative traits such as yield, the most efficient genomic approach is genomic selection (GS, Meuwissen et al. 2001). GS is a method of marker-assisted selection (MAS) which when combined with specific statistical approaches such as BLUP is able to take advantage of the information provided jointly by a large number of markers spread along the whole genome. Advances in genomics also recently made MAS possible for two traits related to yield with simple inheritance, i.e., fruit form (Singh et al. 2013; Ooi et al. 2016) and acidification due to an endogenous lipase (Domonh do et al. 2018); however, this is beyond the scope of the present article, which is dedicated to the genetic improvement of the quantitative components of palm oil yield.

The methodological changes that punctuated the history of oil palm breeding for yield for a century strongly affected the rate of genetic gain. This review article analyzes these changes. First, we present the approaches implemented for the genetic improvement of palm oil yield (mass selection, MRRS, clonal selection, and GS) and the associated statistical

Fig. 1 Breeding and seed production in oil palm. **a** Artificial pollination (CRAPP, Benin). **b** Bunch partitioning into peduncle, spikelets, fruits, and seeds to measure physical characteristics. **c** Soxhlet extractors to measure the percentage of oil in pulp (CRAPP, Benin). **d** Seed garden (CamSeeds, Cameroon)



methods used to estimate the genetic values. Given its expected importance, GS is presented in greater detail. Finally, the fact that the BLUP method, despite its crucial importance for breeding, has yet to be adopted by the whole oil palm breeding community, and because its use becomes unavoidable with GS, we also provide a practical example of its application with R software and an oil palm toy dataset.

Mass selection

The genetic improvement of palm oil production started in the 1920s, in South-East Asia (SEA, Indonesia and Malaysia) and in what was then known as Belgian Congo (Demol 2002; Corley and Tinker 2016, p. 138), and was based on mass selection (candidates selected on their phenotype).

In SEA, the palm oil industry developed from one planting material, four *D* seedlings introduced into Java (Indonesia) in 1848 from an unknown part of Africa (Demol 2002; Corley and Tinker 2016, p. 6). The narrow genetic base followed by several generations of selection led to a relatively homogeneous and inbred breeding population called Deli (Demol 2002; Corley and Tinker 2016, p. 6). The Deli can be further divided in several subpopulations, such as Marihat Baris, Elmina, etc. (Durand-Gasselin et al. 2000; Demol 2002; Corley and Tinker 2016).

In Africa, as the source palms were of *D*, *T*, and *P* types, the breeding approaches differed from those used in SEA (Durand-Gasselin et al. 2000; Corley and Tinker 2016). Breeding was less efficient in Africa, as it was complicated by the segregation of the fruit types in the crosses between the best *T*s (Durand-Gasselin et al. 2000; Corley and Tinker 2016). However, it led to the creation of several breeding populations: La Mé (Côte d'Ivoire), Yangambi (Democratic Republic of Congo), Ekona (Cameroon), WAIFOR (Nigeria), etc. The La Mé population originated from 19 individuals selected from prospections made in the 1920s. The Yangambi population dated from the 1920s and originated from 10 to 20 *T*s, included the Djongo palm which given its exceptional qualities would have finally contributed more than 70% to the Yangambi population (Demol 2002; Cochard 2008; Corley and Tinker 2016).

Also, exchanges of breeding material led to the creation of the AVROS breeding population (Indonesia, Malaysia) from the Djongo.

Mass selection with the early breeding populations had been efficient as some components of OY had a moderate level of narrow-sense heritability h^2 such as M/F (0.53) and BW (0.39) (Corley and Tinker 2016, p. 174,180). However, the other components (BN, F/B, and O/M) had low h^2 (< 0.25). This, and perhaps from knowledge of the advancement of breeding methodology from other crops, prompted

the adoption of the more complex breeding schemes described below.

The breeding populations inherited from this period of mass selection can be classified in two complementary groups (A and B) based on the characteristics of their bunch production. Group A, mostly from SEA (i.e., Deli population) and Angola, although the latter has been of lesser importance, produces a small number of big bunches. Group B, comprising the other African populations (with La Mé and Yangambi currently being the most widely used) and AVROS, produces a large number of small bunches (Meunier and Gascon 1972). The complementarity of the FFB yield component traits in the two groups resulting in hybrid vigor explaining the choice of $A \times B$ cross hybrid breeding approaches.

This period was also marked by a major finding, namely the understanding of the genetic control of the fruit type by a gene, now named *SHELL*, with two codominant alleles Sh^- and Sh^+ (Beirnaert and Vanderweyen 1941). *P* type $Sh^-//Sh^-$ and *D* $Sh^+//Sh^+$ are thus homozygotes and *T* $Sh^+//Sh^-$ heterozygote. The type cultivated in commercial plantations since the 1950s is *T*, as it combines a high M/F with female fertility, and is obtained by the cross $D \times P$. Its use instead of the traditional *D* increased oil palm yield by 30% (Corley and Tinker 2016, p. 7).

Current breeding schemes

The breeding schemes currently applied to improve oil palm yield involve two major improvements over mass selection: they exploit the hybrid vigor for bunch production that appeared in the $A \times B$ crosses, and they enable better estimates of genetic values. These schemes are mainly modified reciprocal recurrent selection (MRRS), which generates sexual crosses (Fig. 1d), which account for the vast majority of oil palm commercial varieties grown in plantations; and clonal selection. They use mating designs, experimental designs and methods of statistical analysis that more efficiently separate the different genetic and environmental effects.

Mating designs and experimental designs

In oil palm MRRS, the selection candidates are evaluated in hybrid crosses obtained according to NCM1 (NCM, North Carolina model) or NCM2 mating designs (Soh 1999). The NCM1 is a hierarchical mating design in which each individual belonging to group B is crossed with a set of different individuals belonging to group A. If individuals in group A can be considered as genetically homogeneous, NCM1 gives satisfactory estimates of the relative genetic or general combining ability values in group B. The NCM2 is a factorial design in which each B individual is crossed with the same set of A individuals (Corley and Tinker 2016, p. 159). This

takes longer as several crosses have to be made per individual in group A but is more suitable than NCM1 when genetic variability among the A individuals is not negligible or when the interactions between parents (i.e., specific combining abilities, SCA) need to be estimated.

Once the crosses or the clones to be evaluated have been obtained, they are planted in field trials, usually according to randomized complete block designs (RCBD). The RCBD used in oil palm breeding usually have 10 to 50 families repeated three to six times in plots each of which contain 12 to 30 palms (Soh et al. 2017, p. 333). Given the low planting density of oil palm (normally 143 individuals per hectare), the trials require a large area (often > 10 ha) whose environmental conditions are consequently subject to some heterogeneity. To better account for this heterogeneity, the complete blocks can be divided into incomplete blocks, i.e., comprising a sample of the evaluated families randomized within the complete blocks (Breure and Verdooren 1995; Soh et al. 2017). Several experimental designs with incomplete blocks are thus commonly used for oil palm, including squared balanced or unbalanced lattices and alpha-plans (Soh et al. 2017, p. 330). The results of evaluations of such trials using RCBDs and lattices have been published for hybrid crosses (Soh et al. 2017, p. 330) and clones (Nouy et al. 2006). In experiments to study the genotype (G) \times environment (E) interaction, the most commonly used design is the split plot. In this case, E is the main treatment (planting density, fertilization, etc.) and G the sub-treatment (parents, hybrids or clones), which facilitates the management of the sub-plots and improves the statistical analysis, as the sub-treatment and the interaction effects are estimated more accurately (Soh et al. 2017, p. 330). For instance, in a trial based on a split plot design with planting density as the main treatment and hybrid crosses as sub-treatment, Rafii et al. (2013) found significant effects of G \times planting density interactions on the average bunch weight.

Modified reciprocal recurrent selection

Principle

Reciprocal recurrent selection (RRS) was defined by Comstock et al. (1949) in maize. It relies on the joint and reciprocal improvement of two heterotic groups. A modified version of reciprocal recurrent selection (MRRS) was adapted for oil palm (Gascon and De Berchoux 1964) and implemented by the IRHO in Côte d'Ivoire (CNRA), Cameroon (IRAD), Benin (CRAPP), and Indonesia (SOCFINDO, IOPRI) (Meunier and Gascon 1972; Corley and Tinker 2016, p. 138; Cochard et al. 2018). In oil palm, MRRS is justified by the fact that in A \times B crosses, the production of bunches is > 25% higher than in the parental populations (Gascon and De Berchoux 1964). This is the result of the negative correlation between BW and BN within each group, and from the

complementarity of groups A and B for these two traits (Table 1). Today, MRRS is used in many countries and although its implementation varies among research centers, it generally follows the scheme described below. However, a number of programs in Malaysia, Indonesia, and Papua New Guinea also practice the modified recurrent selection (MRS) or FIPS (family and individual palm selection) in which *D* and *T* parents for further breeding are recurrently mass selected and the *D* \times *P* progeny testing is done to identify the parents, especially the *Ps*, used for *D* \times *P* seed production (Soh et al. 2017).

One cycle of oil palm MRRS (Fig. 2) starts with selection of candidates from groups A and B, and after evaluation in hybrid progeny tests, the best ones will be selected among them. These candidates will then be used to produce the next generation, which will be used to produce seeds of *T* hybrids and to start a new MRRS cycle (Meunier and Gascon 1972). In more detail, a cycle starts with phenotypic preselection prior to progeny tests. In group A, the individuals are selected based on their own phenotypic value for the traits with the highest heritability (mostly M/F) and on the mean performance of their family (i.e., FIPS). In group B, the female sterility of *P* means they can only be selected based on the mean value of their *T* full-sibs. For the same reason, and to be able to produce the following B generation, *T* individuals are also chosen by FIPS. Second, the combining ability of these individuals in hybrid crosses is evaluated in progeny tests, for the selection of low heritability traits and to finalize the selection of the traits subjected to the first stage of selection. For this purpose, the hybrids crosses are made according to the previously described mating designs, B individuals being crossed with three to five *D* belonging to group A (Soh et al. 2010). These crosses are then evaluated in field trials, during which data are usually recorded from the third year after planting (i.e., at the beginning of production) to the tenth year. A long time is therefore required to obtain the genetic value of the progeny-tested individuals, resulting in long selection cycles lasting around 20 years. The resources required to carry out such long-term evaluations limit the number of individuals that are progeny tested, which results in the erosion of genetic diversity. To address this problem, new germplasms, for example originating from other breeding programs, are introduced (Jacquemard et al. 1997, p. 516).

When analyzing the phenotypic data of the progeny tests, the total genetic value of a hybrid cross is partitioned into the additive value or GCA of its parents or the non-additive or SCA of the cross. The GCA of a parent is the mean value of all the crosses that can be made between this parent and the parents of the other group, expressed as the difference from the mean value of all possible hybrid crosses (Gallais 2011; Corley and Tinker 2016). The SCA of a cross is the difference between the observed value of the cross and the value predicted from the GCA of its parents (Gallais 2011). It represents the interaction between its parents

Table 1 Origin of heterosis in oil palm for bunch yield (figures are indicative)

	Annual number of bunches	Average bunch weight (kg)	Bunch yield (kg/an)
Group A	10	20	200
Group B	20	10	200
A × B hybrid	15	15	225

and usually results from dominance and/or epistatic effects (Stuber and Cockerham 1966; De Souza 1992). It can also result from the multiplicative interaction between two negatively correlated traits as BN and BW for FFB production in oil palm. In this case, SCA may be present even in the absence of non-additive genetic effects (Schnell and Cockerham 1992; Gallais 2011, pp. 68, 71). Finally, the parents with the best GCAs and/or resulting in the crosses with the best SCAs are selected. However, the SCAs for the components of oil palm yield are a much smaller source of variation among the hybrid performances than the GCAs and are estimated with a lower accuracy than the GCAs (Cros 2014). For these reasons, selection is mostly made on the GCAs (Breure and Verdooren 1995; Cros 2014).

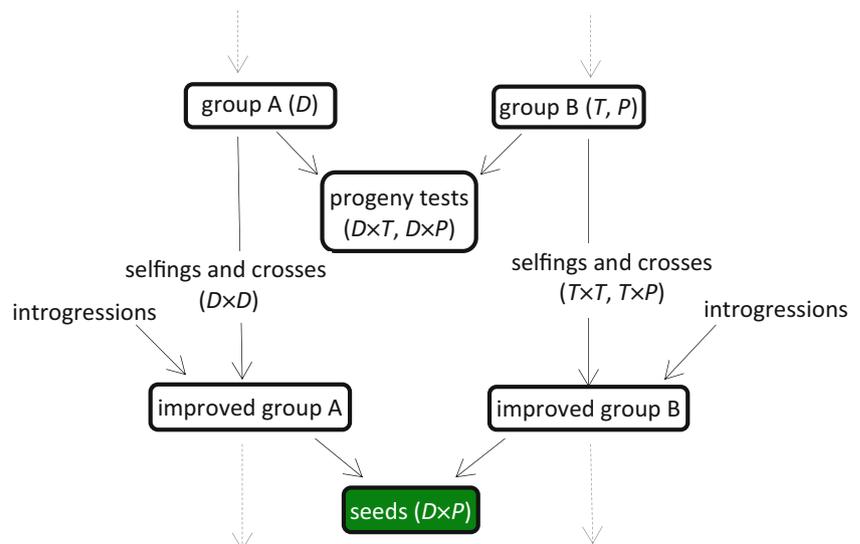
Statistical methods to estimate genetic values

According to the number of published articles, ANOVA is still the most widely used method to estimate GCAs in oil palm, and even to estimate the total genetic value of hybrid crosses without partitioning it into GCAs and SCAs (see for example Breure and Bos 1992; Okwuagwu et al. 2008; Okoye et al. 2009; Junaidah et al. 2011; Noh et al. 2012; Arolu et al. 2016). To estimate the parental GCAs using ANOVA in a hybrid trial set up according to a RCBD, it can be considered that the yield y_{ijk} of cross $A_i \times B_j$

measured in block k is given by the model: $y_{ijk} = \mu + b_k + GCA_i + GCA_j + \varepsilon_{ijk}$, where μ is the phenotypic mean of the trial, b_k the effect of block k , GCA_i and GCA_j the parental GCAs, and ε_{ijk} the error associated with the k^{th} replicate of the cross (Breure and Verdooren 1995), with $y_{ijk} \sim N(E(y_{ijk}), \sigma^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$. The solutions of the model (i.e., the least square means), and in particular the parental GCAs, are obtained by the ordinary least squares method. The SCAs are then obtained by subtracting the cross values expected from the parental GCAs from the mean cross values observed in the trial. ANOVA is useful for complete or balanced experimental designs and mating designs.

However, it is also possible to estimate the genetic values with the BLUP method, which is the standard approach for analyzing linear mixed models. BLUP was developed several decades ago to analyze highly unbalanced datasets in cattle breeding. Today, it is widely used to estimate genetic effects in animals (Mrode 2005) and in plants (Piepho et al. 2008). BLUP has the following advantages (Soh 1999): it is useful in analyzing unbalanced mating designs or experimental designs, and it makes it possible to consider a large number of trials at the same time, even without control families, and to account for covariances when modeling, for example, the relationships among individuals, competition effects, or spatial heterogeneity. Surprisingly, in oil palm, it has only been used to estimate genetic values for yield components by a very limited number of research groups (Soh 1994; Purba et al. 2001; Cros et al. 2015b). However, oil palm progeny tests are often carried out with complex and unbalanced designs, with a varying number of crosses per parent, crosses evaluated in several trials planted in different years, varying numbers of replicates and individual palms per cross, etc. The mating design is also sometimes not connected, i.e., that within a parental group, some parents are not connected (directly or indirectly) to the others by the same partners that belong to the other group, even though this can bias or make the GCA of some parents impossible to estimate (Breure and Verdooren 1995; Soh et al. 2017).

Fig. 2 Scheme of one cycle of modified reciprocal recurrent selection applied to oil palm. *D* *dura*, *T* *tenera*, *P* *pisifera*, green: commercial seeds



Several studies have also shown that in such complex situations, ANOVA was less efficient than BLUP in estimating the variances and/or the effects in the model (White and Hodge 1989; de Carvalho et al. 2008, p. 220; Piepho et al. 2008; Hu 2015). In addition, the pedigree of the oil palm breeding populations over several generations is generally known (Cros et al. 2014; Corley and Tinker 2016, pp. 138–148), and the relationships among selection candidates is useful information that can be included in the linear mixed model in order to more accurately estimate the genetic parameters and the genetic values.

In the case of hybrid crosses between two parental populations A and B, the linear mixed model used to estimate the parental GCAs and the cross SCA is

$$y = X\beta + Z_1u_A + Z_2u_B + Z_3u_{AB} + \varepsilon$$

with y the vector of observed phenotypes, β the vector of fixed effects, $u_A \sim N(0, 0.5A_A\sigma_{aA}^2)$, and $u_B \sim N(0, 0.5A_B\sigma_{aB}^2)$ the vectors of the GCAs of parents of groups A and B (random effects), respectively, and $u_{AB} \sim N(0, 0.25D_{AB}\sigma_{ascAB}^2)$ the vector of cross SCA, corresponding here to the dominance effects (random). X , Z_1 , Z_2 , and Z_3 are, respectively, the incidence matrices associated to β , u_A , u_B , and u_{AB} . $\varepsilon \sim N(0, I\sigma_\varepsilon^2)$ is the vector of residual effects and I is the identity matrix (in this example, residuals are assumed to be independent). $0.5A_A\sigma_{aA}^2$, $0.5A_B\sigma_{aB}^2$, and $0.25D_{AB}\sigma_{ascAB}^2$ are the variance-covariance matrices associated with u_A , u_B , and u_{AB} , respectively. A_A and A_B are the matrices containing the values of additive relationships calculated with the pedigree of the A and B individuals, respectively, and D_{AB} is the matrix of dominance relationships between the crosses and is obtained by the Kronecker product between A_A and A_B . σ_{aA}^2 and σ_{aB}^2 are the additive genetic variances of groups A and B, respectively, and σ_{ascAB}^2 is the dominance genetic variance of the crosses. The BLUP approach starts with estimation of the variances σ_{aA}^2 , σ_{aB}^2 , σ_{ascAB}^2 , and σ_ε^2 . The most widely used method for this purpose is restricted maximum likelihood (REML) (Xavier et al. 2016). Various algorithms have been developed to estimate the variance components with REML. The two main ones are the expectation-maximization algorithm (EM), which relies on the iterative updating of the residuals, variances, and regression coefficients of fixed and random effects (Dempster et al. 1977), and the average-information algorithm, which relies on the creation of a gradient based on the mean of the expected and observed information (Gilmour et al. 1995). Second, the variances are used in the mixed-model equations of Henderson, which give the model solutions, i.e., the vectors \hat{u}_A , \hat{u}_B , and \hat{u}_{AB} for the genetic effects and the vector $\hat{\beta}$ for the fixed effects (Covarrubias-Pazarán 2016). The solutions are named best linear unbiased estimators (BLUE), or solutions of the generalized least squares, for the fixed effects, and best linear unbiased predictors

(BLUP) for the random effects (Mrode 2005 p. 39–42). The method also makes it possible to estimate the accuracy of the BLUPs, i.e., their correlation with the true genetic values that the model estimates. The accuracies are given by a theoretical formula using the diagonal of the variance-covariance matrix of the random effect considered and the prediction variance errors (PEVs) associated with the BLUPs, which are easily obtained from the analysis. Thus, with the model presented here, the accuracy of the GCA of parent A_i is

$$r_{u_{A_i}, \hat{u}_{A_i}} = \sqrt{1 - \frac{\text{PEV}_{u_{A_i}}}{0.5(1+F_{A_i})\sigma_{aA}^2 \hat{u}_{A_i}}}$$

with $0.5(1+F_{A_i})\sigma_{aA}^2$ the i th element of the diagonal of the variance-covariance matrix of u_A , and F_{A_i} the inbreeding coefficient of A_i (Cros 2014). The application of this formula in oil palm showed that for the yield components, the hybrid progeny tests gave highly accurate GCAs, reaching on average 0.87 in group A and 0.91 in group B (Cros 2014).

To promote the adoption of this method by the largest number of geneticists, in particular in the oil palm breeding community, in Appendix, we provide a practical example of the estimation of the BLUP value of parents of oil palm hybrids using R software (R Core Team 2017).

Clonal selection

The main use of clonal selection in oil palm is cloning the best T hybrid individuals. For this purpose, the T with the best phenotypes are chosen within the best crosses available in the MRRS program and are evaluated in clonal trials (Corley and Tinker 2016, pp. 216–220). The interest of this method is based on oil palm heterozygosity, which generates genetic variability within the hybrid crosses, allowing selection of the best T individuals to be used as ortets (source plants for cloning). The clones have the potential to further increase oil palm yield by 20% to 30% compared to sexual crosses (Corley and Law 1997), and increases in yield of 13% (Nouy et al. 2006) and 18% (Soh et al. 2003a) have been empirically observed. One difficulty in clonal selection is to accurately estimate the genetic value of the hybrid individuals from their own phenotypic records, given the micro-environmental effects that are hard to control and are confounded with individual genetic values. This accuracy can be measured by the broad-sense heritability H^2 computed at the individual level. Soh et al. (2003b), Nouy et al. (2006), and Potier et al. (2006) showed that H^2 ranged from 0 to 0.84 among yield components. In these conditions, it is possible to select ortets based on their phenotype for some traits, such as O/M, but not for all yield components. Clonal field trials are thus required to finalize the

evaluation of the ortets selected based on the traits with the highest H^2 . These trials allow a highly reliable selection of ortets but lengthen the selection process by at least 10 years, corresponding to the time required to produce the clones from explants and to carry out the trial, thus allowing improved hybrids to catch up and reduce the advantage of clones.

Oil palm cloning has been slowed down by the appearance of abnormal floral morphogenesis in the field. The abnormal ramets, or mantled variants, produce abnormal flowers and fruits and bunch failure, leading to sterile palms (Soh et al. 2017, p. 172). The epigenetic molecular mechanism that causes this abnormality was recently elucidated. The mantled variants were shown to result from hypomethylation during tissue culture of the Karma retrotransposon, located in the intron of the *DEFICIENS* gene. This altered its splicing and made it produce an additional transcript associated with the mantled phenotype (Ong-Abdullah et al. 2015; Soh et al. 2017, p. 207). The understanding of this mechanism opens the way for the development of a molecular kit that will allow the early detection and elimination of abnormal ramets, thus boosting interest in oil palm cloning. Research is also underway to broaden the range of genotypes in which tissue culture is efficient (Soh et al. 2017). In addition, cloning opens the way for the production of genetically engineered palms. Indeed, tissue culture is an appropriate way to regenerate genetically modified tissue, and several genetic transformation methods have been successfully applied in oil palm (biolistic, transformation with *Agrobacterium*, and microinjection) (Masani et al. 2018).

Advantages and drawbacks

The current breeding schemes have the advantage of accurately estimating the genetic values, thereby enabling efficient selection, which, in turn, has enabled the significant genetic progress achieved so far. However, the schemes also have two drawbacks resulting from the difficulties involved in phenotyping. First, as mentioned above, the breeding cycle to produce a new variety is long, around 20 years, whereas oil palm reaches sexual maturity relatively quickly (at 3 or 4 years old). The length of the cycle is mostly due to the phase of evaluation in progeny tests, as a long time is required to make the crosses, obtain the plants, and above all, to carry out the field trial. Second, these schemes have low selection intensity, with—for example—fewer than 200 selection candidates progeny tested per population and cycle. The first stage of selection before the field trials (progeny tests or clonal trials) based on the phenotypic values for the most heritable traits seems to compensate for the reduced number of parents or clones evaluated, but this is not optimal. Indeed, the first stage of selection is made on a small number of traits and its accuracy is lower than selection based on progeny tests or clonal trials. Consequently, the individuals that would be the best considering their genetic value over all the yield components

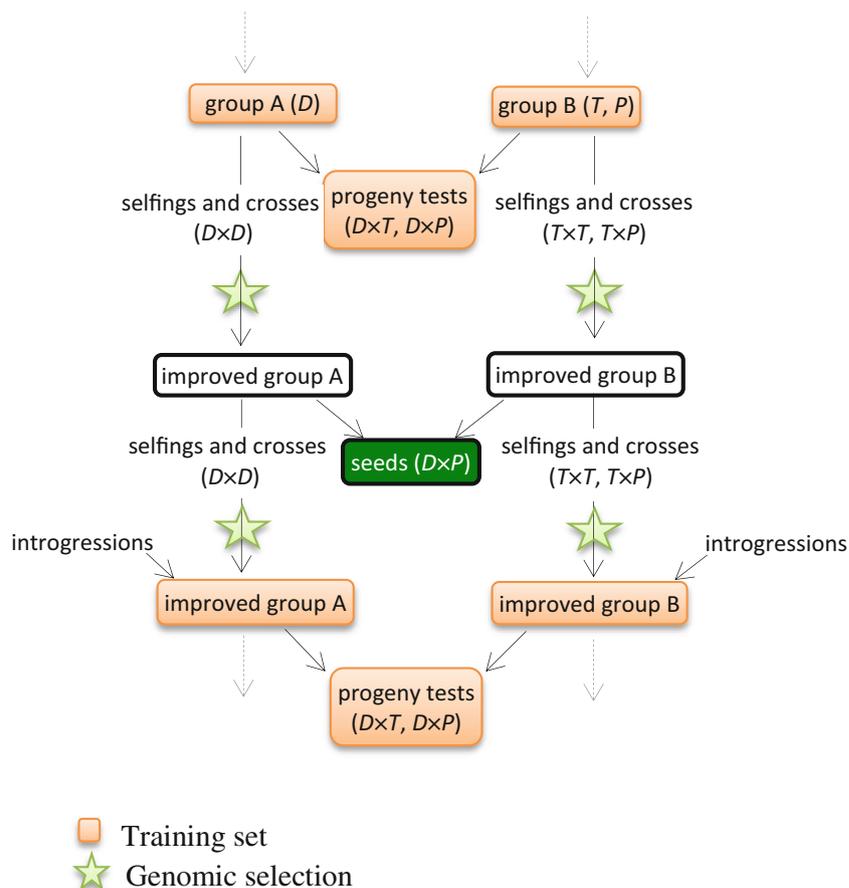
may be discarded before the field trials because they do not have the best phenotypic value for the trait or the few traits used in the first stage of selection. This even led to questioning the relevance of the first selection stage prior to field trials. For clonal selection, the possibility of randomly choosing the ortets before evaluating them in clonal trials has thus been considered by several authors (Corley and Tinker 2016, p. 216). However, to be efficient, this method would require exploring a large part of the genetic variability of the hybrid crosses where the ortets would be chosen, i.e., evaluating a large number of candidate ortets in clonal trials, which is not feasible in practice. New methods are therefore required to optimize the current breeding schemes.

Genomic selection

The first saturated genetic maps were produced at the end of the 1980s. They made it possible to detect QTLs (quantitative trait loci), leading to the idea of MAS. MAS has the potential to increase selection intensity and shorten the breeding cycles (Muranty et al. 2014). Many QTLs related to oil palm yield have been identified (see for example Billotte et al. (2010), Pootakham et al. (2015), Tisé et al. (2015), Ting et al. (2018)). However, for complex traits such as yield that are under the control of a large number of genes with small effects, the efficiency of the approach is limited, in particular in the case of small population size (Muranty et al. 2014), because it overestimates the effect of the strong QTLs and fails to exploit weak QTLs, as their effect does not appear to be significant (Muranty et al. 2014). A more efficient approach, genomic selection (GS), was consequently developed (Meuwissen et al. 2001). Its practical implementation was made possible by progress in genomics, in particular in next-generation sequencing (NGS) and high throughput genotyping. Today, GS is used in animal breeding, particularly in dairy cattle, where it has doubled the rate of the genetic progress (Wiggins et al. 2017). In plants, it is progressively being incorporated in breeding schemes, and it is expected to significantly increase their efficiency (Varshney et al. 2017).

In oil palm, the use of GS to select the parents of the hybrid crosses for yield traits has already been investigated in several studies. They evaluated its ability to reduce the length of the breeding cycles, by avoiding field trials in some cycles, and to increase selection intensity, by the application of selection to a larger number of candidates than with the current method (Fig. 3). The results are promising and are detailed below. So far, no study has been published regarding the use of GS to select ortets, but its potential is likely also high, as suggested by the positive results obtained in other species, and in particular in other perennial tropical crops like eucalyptus (Durán et al. 2017) and rubber tree (Cros et al. [Under review](#)).

Fig. 3 Possible scheme of genomic modified reciprocal recurrent selection applied in large populations of seedlings to increase selection intensity (cycles 1 and 2) and shorten breeding cycles (cycle 2) of oil palm. *D*: *dura*, *T*: *tenera*, *P*: *pisifera*, green: commercial seeds



Principle

GS is MAS for quantitative traits using high-density molecular markers covering the whole genome, in order to have every QTL in linkage disequilibrium with at least one marker. What mainly differentiates it from QTL-based MAS is the joint exploitation of strong QTLs (i.e., whose effect would be shown to be significant in a QTL analysis) and of weak QTLs (not significant). Its goal is to predict the genetic value of selection candidates, usually with no data on their performance (i.e., depending on the breeding situation concerned, with no known phenotype or no progeny tests). For this purpose, GS uses the genotypic and phenotypic data of a population called the training (or calibration) population and a linear mixed model that can predict the additive genetic value (GEBV, genomic estimated breeding values) or the total genetic value (i.e., including the non-additive effects) of the selection candidates (Heffner et al. 2009). GS therefore has the potential to reduce phenotyping, thus making it possible to shorten the breeding cycle and/or to increase selection intensity.

The efficiency of GS is assessed by computing its selection accuracy (r_{GS}), i.e., the correlation between the genetic value estimated with the genomic model (GEGV) and the true genetic value (TGV) in a set of individuals used as the validation population. However, in empirical studies, the true genetic value is

unknown, and the genetic value estimated with the genomic model is therefore correlated with an estimate of the true genetic value (EGV), obtained with the phenotypic data available on the validation individuals, i.e., their own phenotypic records or the phenotypes of their progenies. This correlation is named prediction accuracy. The difference between selection accuracy and prediction accuracy depends on the reliability of the EGV (Lorenz et al. 2011, p.94). GS accuracy is crucial to evaluate the potential of GS as it is directly related to the rate of the genetic progress, or rate of selection response $R = r_{GS} \times i \times \sigma_g / L$, with σ_g the genetic variance and L the generation interval (Falconer and Mackay 1996). However, a comprehensive comparison of GS and conventional selection requires considering their respective selection accuracy, selection intensity, and generation interval. Indeed, even in a situation where GS accuracy would be lower than the accuracy of the conventional phenotypic evaluations, GS can still increase R if it allows a sufficient decrease in the generation interval and/or increase in selection intensity.

GS accuracy is affected by several parameters, including marker type and density, distribution of QTL effects, linkage disequilibrium between markers and QTLs, the size of the training population, and the relationship between the training and selection populations, trait heritability, and statistical methods of prediction (Lorenz et al. 2011; Grattapaglia 2014). In practice,

GS accuracy is usually estimated by cross-validation at a single experimental site (Cros et al. 2015b; Kwong et al. 2017a, b) or by between-site validation (Cros et al. 2017). However, single-site cross-validations may overestimate accuracy, and it is therefore preferable to have at least two sites to evaluate GS (Lorenz et al. 2011, p.94).

Molecular data

GS generally uses single nucleotide polymorphism markers (SNPs). They are abundant on the whole genome, have a low mutation rate (Oraguzie et al. 2007, p. 41), and can easily be genotyped at reasonable cost. In oil palm, given the molecular resources available at the time, the first empirical studies were made with microsatellites (SSR, simple sequence repeats) (Cros et al. 2015b; Marchal et al. 2016). However, GS studies in this species now use SNPs from genotyping by sequencing (GBS) (Cros et al. 2017) or SNP arrays (Kwong et al. 2016, 2017a, b; Ithnin et al. 2017). This allowed reaching higher densities, which contributed to achieve higher accuracies. Thus, Kwong et al. (2017b) using 135 SSRs obtained mean GS prediction accuracies of 0.21 over palm oil yield components, against 0.31 with 200 K SNPs.

GS accuracy normally increases with the number of markers until it reaches a plateau (de los Campos et al. 2013, p. 339; Cros 2014, p. 40). In oil palm, the effect of marker density on the GS accuracy for yield components has been evaluated in three studies. When predicting the performance of unevaluated hybrids, GS accuracy started plateauing with 500 and 2000 SNPs in Cros et al. (2017) and between 200 and 400 SNPs in Kwong et al. (2017a), depending on the trait. The two studies did not consider the same populations, but the smaller number of SNPs required in Kwong et al. (2017a) likely resulted from the fact that the SNPs were chosen based on the association scores estimated in a genome-wide association study, and not randomly, as in Cros et al. (2017). When predicting the GCA of progeny-tested individuals, Marchal et al. (2016) showed that GS accuracy plateaued with 160 SSRs in group A and 90 SSRs in group B. The marker density required to reach the maximum GS accuracy therefore varies depending on the type of marker, the marker sampling method, the trait, and the population. However, the marker density needed in oil palm is lower than is generally the case in other species due to the high rate of inbreeding in oil palm breeding populations, i.e., to their small effective size (Cros et al. 2014).

Genotyping generates missing data. There are very few missing data with SNP arrays (< 1% in Kwong et al. (2016)) and SSRs (< 3% in Cros et al. (2015b)), but they can reach significant proportions with GBS (13.2% in Cros et al. (2017)). The GS statistical models cannot deal with missing molecular data, which therefore have to be imputed. This consists in replacing them by the most likely genotype. In practice, the imputation method is likely of no importance when the percentage of missing data is

low. In this case, the missing data can be replaced by the genotype with the highest frequency for the marker considered in the population concerned, as in Kwong et al. (2017a). With more missing data, more sophisticated imputation approaches are recommended. Many methods are available for this purpose (Wang et al. 2016). Currently, only the BEAGLE software (Browning and Browning 2007) has been used to impute missing molecular data in GS studies on oil palm. Cros et al. (2017) showed that taking pedigree information into account for imputation made BEAGLE more efficient. However, they also noted that, for a given number of markers, using those with the lowest percentage of missing data resulted in higher GS accuracy than using random markers, which suggests that imputation could be improved.

Training and application populations

GS accuracy normally increases with the size of the training population (Lorenz et al. 2011; Grattapaglia 2014) and with the relationship between training and application individuals (Pszczola et al. 2012). In oil palm, GS accuracy was observed empirically to be strongly affected by the relationship between training and application individuals (Cros et al. 2015b), suggesting that the use of GS in full-sibs or progenies of the training individuals would maximize accuracy. To increase the size of the training set, it is possible to aggregate data from consecutive breeding cycles. Simulations in oil palm showed that using data from two cycles increased the per cycle response to selection by more than 10%, mainly as a result of higher selection accuracy (Cros et al. 2018). Although this aggregation of data reduces the relationship between training and application populations, this is more than counterbalanced by the doubling of the training population.

Several strategies can be used to optimize the training and application populations. For instance, the CDmean criterion, derived from the generalized coefficient of determination, can optimize the sampling of individuals that have to be phenotyped among a set of genotyped individuals, in order to form the training population (Rincent et al. 2012). In oil palm, the CDmean proved to be efficient for GS as it maximizes its accuracy (Cros et al. 2015b). However, further improvements are possible: for example, another optimization criterion recently developed to define training populations, CDpop, could be more efficient for oil palm as it is specific to highly structured populations (Rincent et al. 2017).

Models and statistical methods for genomic predictions

Genomic predictions are made with frequentist and Bayesian statistical approaches (Varshney et al. 2017). Some methods estimate an effect associated with each marker, while other methods give the genetic values directly without estimating

marker effects. Genomic predictions exploit two types of information, the relationship between training and application populations, and the linkage disequilibrium between markers and QTLs (Varshney et al. 2017).

In methods that estimate marker effects, the base (i.e., purely additive) genomic linear mixed model is of the form: $y = X\beta + Zm + e$, where y is the vector of data records ($n_{\text{ind}} \times 1$), β the vector of fixed effects (mean, trials, blocks, etc.) associated with incidence matrix X , m the vector containing the substitution effect of each SNP ($n_{\text{SNP}} \times 1$) with incidence matrix Z ($n_{\text{ind}} \times n_{\text{SNP}}$) containing the molecular data coded in the number of copies of the most frequent allele (0, 1 or 2), e the vector of residuals ($n_{\text{ind}} \times 1$), n_{ind} the number of individuals in the training population, and n_{SNP} the number of SNPs (Soh et al. 2017, p. 156). The effects m and e are random. The GEBV of selection candidate i is given by summing the SNP effects over the whole genome according to the formula: $\text{GEBV}_i = \sum_{j=1}^{n_{\text{SNP}}} Z_{ij} \hat{m}_j$, with \hat{m}_j the estimated effect of SNP j . Depending on the way the marker genetic variance (σ_m^2) is treated, two types of methods can be distinguished (Soh et al. 2017, p. 156). First, some methods consider that marker effects are sampled according to a normal distribution with a variance common to all markers, which is relevant for traits following the infinitesimal model. This is the case of random regression BLUP (RR-BLUP) (Meuwissen et al. 2001) and Bayesian random regression (BRR) (Pérez et al. 2010). Second, as the genetic determinism of some quantitative traits may include loci with strong effects, other methods such as Bayes A, Bayes B (Meuwissen et al. 2001), Bayes C π , Bayes D π (Habier et al. 2011), and Bayesian LASSO (De Los Campos et al. 2009) attribute marker-specific genetic variances.

The most widely used method to estimate GEBV directly is the genomic best linear unbiased predictor (GBLUP). The basic difference between GBLUP and conventional BLUP presented above is the use of genomic (instead of genealogic) information to compute the relationship matrix, called the G matrix in GBLUP. The G matrix has the advantage of accounting for the random sampling of alleles at meiosis (Mendelian sampling) and thus gives realized relationships, making it possible to obtain the GEBV of unevaluated individuals. Also, genomic data are not affected by pedigree errors in the families used in the breeding program. By contrast, the pedigree-based A matrix gives expected relationships (Habier et al. 2007; VanRaden 2007), and therefore does not differentiate between individuals within families, cannot capture relationships that do not appear in the pedigree records, and gives erroneous values in the case of illegitimacy. The base model used with GBLUP is $y = X\beta + g + e$, with g the vector ($n_{\text{ind}} \times 1$) of GEBVs following $N(0, G\sigma_g^2)$, σ_g^2 the additive variance, and G ($n_{\text{ind}} \times n_{\text{ind}}$) the genomic relationships matrix. With SNP markers, the G matrix is usually computed according

to VanRaden (2007). GBLUP is equivalent to RR-BLUP under the assumption of normality of marker effects and has the advantage of being simple to implement with existing software and of having a reasonable computation time.

Various modeling approaches have been used for genomic predictions in oil palm. The base GS models described above were used in each parental group separately, with data records consisting in parental performances in crosses with the other group, i.e., GCAs (Cros et al. 2015b) or testcross phenotypic means (Wong and Bernardo 2008), and parent genotypes. Ithnin et al. (2017) and Kwong et al. (2017b) applied similar models but used parental phenotypes as data records. They obtained low to intermediate GS prediction accuracies but, as parental phenotypes may not reflect performance in hybrid crosses due to gene-frequency differences between parental populations and non-additive effects (Wei et al. 1991; Baumung et al. 1997; Vitezica et al. 2016), the relevancy of such accuracies for hybrid breeding is questionable. Kwong et al. (2016) studied GS with a population consisting in a mixture of Deli, group B, and hybrid individuals. They obtained a prediction accuracy of 0.65, which could have possibly been improved by the use of a model designed to jointly consider parental and hybrid data, like in Vitezica et al. (2016). Accuracy of GS could also be improved by a single-step GBLUP (ssGBLUP) which blends realized relationship of genotyped individuals with the genealogical relationship of non-genotyped individuals to calculate GEBV. This increases the size of the training set by taking into account ungenotyped individuals for which phenotypes are available. In oil palm, this could be used to include in the training set phenotyped individuals for which DNA can no longer be obtained, such as individuals evaluated in past progeny tests. In eucalyptus, using additional phenotypic information from non-genotyped individuals thus increased GS prediction accuracies by up to 75% (Cappa et al. 2019). Other studies used the conventional MRRS model replacing genealogical relationship matrices by genomic matrices to jointly predict the GEBV of A and B candidates (Cros et al. 2015a, 2017, 2018; Marchal et al. 2016). In order to increase the training size, this method was adapted to include molecular data of individual hybrids, taking into account the parental origin of marker alleles (Cros et al. 2015a). This gave the highest selection accuracies for unevaluated parents, and thus proved to be more efficient than using only parental genotypes to train the model. Kwong et al. (2017a) also used molecular data of individual hybrids but did not consider the parental origin of alleles. So far, the usefulness of modeling the parental origin of marker alleles in oil palm hybrid genotypes has not been investigated. Further studies thus remain necessary to identify the optimal prediction model, in particular depending on the nature of the training data.

In addition, a wide range of statistical methods has been applied to analyze these models, and comparisons showed that

they did not significantly affect the accuracy of GS (Cros et al. 2015b; Kwong et al. 2017b; Ithnin et al. 2017). This suggests that the components of palm oil yield are highly polygenic and follow the infinitesimal model.

Information captured by markers

Without optimizing the training and validation populations, prediction accuracies ranging from 0.14 and 0.73 were obtained for various yield components, confirming the ability of GS models to predict the genetic value of unevaluated selection candidates (Cros et al. 2017; Kwong et al. 2017a, b). In particular, for five yield components (FFB, O/M, BN, BW, and M/F), the GS model predicted the performance of unevaluated hybrid crosses with higher accuracy than a control model using pedigree data instead of markers (Cros et al. 2017). This showed the ability of GS to capture genetic differences within full-sib families (i.e., the Mendelian segregation term) in addition to genetic differences between families, enabling the selection of the best individuals within the best families, as currently done among the individuals that are progeny tested. The same conclusion was reached in Kwong et al. (2017b), where GS prediction accuracies above zero, ranging from 0.18 to 0.47, were obtained in a GS evaluation considering a single full-sib family. Similarly, Cros et al. (2015b) obtained GS prediction accuracies above 0.5 within full-sib families. However, the latter study also showed that GS could also, depending on trait and population, fail to capture Mendelian segregation. In this case, GS predictions only revealed, at the best, between-family differences.

Annual genetic progress

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

In another simulation study, Cros et al. (2015a) compared conventional MRRS and GS over four cycles. With GS, each

cycle including hybrid progeny tests was used to train a model applied to make a selection among unevaluated individuals of the same cycle (i.e., sibs of the evaluated individuals) and/or of the following generations. The effect on the annual selection response of the following parameters was quantified: frequency of progeny tests (from model training only in first cycle to training in every cycle), the number of GS candidates (120 and 300), and GS strategy (genotyping limited to the parents of the calibration hybrids [RRGS_PAR] or also genotyping hybrid individuals [RRGS_HYB]). The authors showed that GS can increase annual genetic progress by reducing the generation interval and by increasing the selection intensity, despite the fact that GS accuracy for unevaluated hybrid parents is lower than the accuracy of progeny tested parents. Among the strategies evaluated, RRGS_HYB with the genotyping of 1700 hybrid individuals, model training only in the first generation, and 300 selection candidates per population and generation was the most efficient, leading to 72% higher annual genetic progress than MRRS. Additionally, RRGS_PAR with model training every two generations and 300 selection candidates was shown to be an interesting alternative as although its genetic progress was lower (46% higher than MRRS), it had a lower variability of genetic progress, reduced cost, and slower increase in inbreeding over cycles in the parental populations compared to RRGS_HYB. The authors later studied the effect of aggregating the data of two consecutive cycles to train the RRGS_PAR model and showed that this increased the selection accuracy, leading to an annual genetic progress 37.6% to 57.5% higher than MRRS, depending on the number of GS candidates (Cros et al. 2018).

These simulation results promise a revolution in the genetic improvement of oil palm yield. However, this needs to be put into perspective by the empirical studies that even if they showed that GS accuracies could be high, also revealed that GS was not efficient for all yield components. Indeed, for some traits, the GS model did not predict the genetic value of unevaluated individuals better than a control model using pedigree data instead of markers (Cros et al. 2015b, 2017). Yet, the simulations showed that the main advantage of GS was its ability to shorten the breeding cycles by avoiding field evaluations in some cycles, and this is only possible if GS is efficient for all the yield components that are currently the subject of phenotypic selection. Otherwise, the progeny tests remain necessary in all breeding cycles. Therefore, the practical application currently envisaged to start implementing GS in oil palm is a two-stage scheme, with an initial stage of genomic selection prior to progeny tests. This would be better than the current first stage of phenotypic selection for two reasons. First, the number of yield components for which GS is efficient is greater than the number of traits currently subjected to phenotypic preselection. Second, the current selection prior to progeny tests is made on the parental phenotypes, even though, as already mentioned, they may be poor indicators of performance in hybrid crosses. By contrast, this would not be a problem for genomic predictions obtained with a model

Table S1 Incomplete NCM2 mating design (the asterisks represent the number of crosses)

<i>Pisifera dura</i>	B3	B5	B7	B9
A5	***	***		
A6		***	***	
A7			***	***
A8	***			***

calibrated on hybrid phenotypes. The potential of genomic preselection was quantified based on the GS accuracies empirically obtained by between-site validation for bunch production, a trait which is normally not subjected to phenotypic selection prior to progeny tests in the current schemes (Cros et al. 2017), and the study showed that this would increase the performance of the selected hybrids by more than 10% compared to a method without preselection, thanks to higher selection intensity.

To be applied in practice, GS must also result in annual genetic progress per unit cost higher than current selection methods. Although GS generates additional costs related to genotyping, these costs are low in comparison to the cost of phenotyping. Thus, Jacob et al. (2017) indicated that, even assuming a genotyping cost per sample as high as 300€, which seems to be the maximum possible price for a 300 K SNP array, the ratio of genotyping/phenotyping costs lays below 1/20. In addition, these extra costs could possibly be offset by a reduction in phenotyping costs, when it is possible to manage without some field evaluations. In this case, Wong and Bernardo (2008) found that with a

Table S2 Additive genetic values of progeny tested individuals

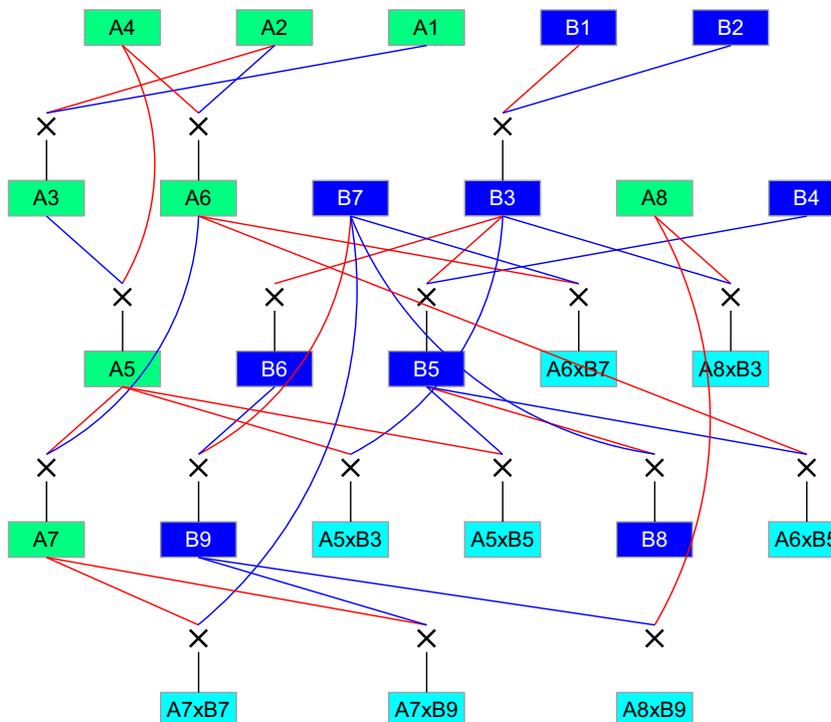
	GCA _s	Standard errors
u_{A_5}	-1.44	6.35
u_{A_6}	6.05	6.39
u_{A_7}	12.64	6.45
u_{A_8}	-3.35	6.35
u_{B_3}	10.60	6.48
u_{B_5}	3.96	6.45
u_{B_7}	-11.53	6.43
u_{B_9}	-1.91	6.49

genotyping cost of US\$0.15 per datapoint, corresponding to genotyping prices for SNPs, the cost per genetic progress unit was 35% to 65% lower with GS than with conventional selection.

Conclusions

The history of the genetic improvement of oil palm was marked by three disruptive improvements that accelerated the rate of the genetic progress: (1) understanding the heredity of the fruit form, which led to the replacement of *D* by *T* in plantations; (2) the discovery of hybrid vigor in bunch production which led to the adoption of hybrid cultivars and to the replacement of mass selection by MRRS; and (3) clonal selection, exploiting intra-hybrid genetic variability. Today, GS appears to be a new

Fig. S1 Pedigree of the example population. Green: individuals from group A; blue: individuals from group B; turquoise: A × B hybrid crosses



$u_A \sim N(0, 0.5A_A\sigma_{a_A}^2)$, $u_B \sim N(0, 0.5A_B\sigma_{a_B}^2)$ and for example, for the eight individuals in the pedigree of group A, the coancestry matrix

$$0.5A_A = \begin{pmatrix} 0.5 & 0 & 0.25 & 0 & 0.125 & 0 & 0.063 & 0 \\ 0 & 0.5 & 0.25 & 0 & 0.125 & 0.25 & 0.188 & 0 \\ 0.25 & 0.25 & 0.5 & 0 & 0.25 & 0.125 & 0.188 & 0 \\ 0 & 0 & 0 & 0.5 & 0.25 & 0.25 & 0.25 & 0 \\ 0.125 & 0.125 & 0.25 & 0.25 & 0.5 & 0.188 & 0.344 & 0 \\ 0 & 0.25 & 0.125 & 0.25 & 0.188 & 0.5 & 0.344 & 0 \\ 0.063 & 0.188 & 0.188 & 0.25 & 0.344 & 0.344 & 0.594 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 \end{pmatrix}$$

The estimates of the variance components were obtained from the syntax:

```
remlf90(fixed = RENDEMENT ~ REP, generic = list(parent_A =
list(Z.mat_A, A.mat_A), parent_B = list(Z.mat_B, A.mat_B)), data =
yield_data)
```

where `remlf90` is the function that analyzes the linear mixed model using the REML, `fixed` is the argument representing the fixed effects (here, replicates), `generic` the argument representing the random genetic effects (GCAs) and indicating for each the associated incidence and variance-covariance matrices (`parent_A` and `parent_B` are the columns in the table `yield_data`). The objects `Z.mat_A` and `Z.mat_B` are the incidence matrices Z_1 and Z_2 , respectively. The objects `A.mat_A` and `A.mat_B` are the matrices $0.5A_A$ and $0.5A_B$ generated by the function `kinship` (package `kinship2`) that computes the genealogical coancestry coefficients between the individuals in the pedigree.

The analysis gives the following variance estimates (\pm standard error): $\sigma_{a_A}^2 = 192.15 \pm 164.58$, $\sigma_{a_B}^2 = 195.36 \pm 164.51$, $\sigma_\varepsilon^2 = 7.32 \pm 2.68$, and the solutions for the block effects (BLUE): $\beta_1 = 6.85 \pm 8.98$, $\beta_2 = 5.78 \pm 8.98$, $\beta_3 = 6.47 \pm 8.98$. The solutions for the GCAs (BLUP) are given in Table S2. According to the parental GCAs, the best possible cross would have been $A_7 \times B_3$, with an expected yield of 29.60 ($\beta + u_{A_7} + u_{B_3}$), while the best cross in the trial was $A_7 \times B_9$, with an expected yield of 20.91 (and a mean observed yield of 21.98).

References

Arolo IW, Rafii MY, Marjuni M et al (2016) Genetic variability analysis and selection of pisifera palms for commercial production of high yielding and dwarf oil palm planting materials. *Ind Crop Prod* 90: 135–141. <https://doi.org/10.1016/j.indcrop.2016.06.006>

Baumung R, Sölkner J, Essl A (1997) Correlation between purebred and crossbred performance under a two-locus model with additive by additive interaction. *J Anim Breed Genet* 114:89–98. <https://doi.org/10.1111/j.1439-0388.1997.tb00496.x>

Beirmaert A, Vanderweyen R (1941) Contribution à l'étude génétique et biométrique des variétés d'*Elaeis guineensis* Jacq. *Publ Inst Nat Etude Agron Congo Belge Ser Sci* 27:1–101

Billotte N, Jourjon M, Marseillac N et al (2010) QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis* Jacq.). *Theor Appl Genet* 120:1673–1687

Breure C, Bos I (1992) Development of elite families in oil palm (*Elaeis guineensis* Jacq.). *Euphytica* 64:99–112

Breure CJ, Verdooren LR (1995) Guidelines for testing and selecting parent palms in oil palm, practical aspects and statistical methods. *ASD Oil Palm Pap* 9:68

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81: 1084–1097

Butler D, Cullis BR, Gilmour A, Gogel B (2009) ASReml-R reference manual. State Qld Dep Prim Ind Fish Brisb, Brisbane City 398 p

Cappa EP, de Lima BM, da Silva-Junior OB et al (2019) Improving genomic prediction of growth and wood traits in Eucalyptus using phenotypes from non-genotyped trees by single-step GBLUP. *Plant Sci*:284, 9–215

Cochard B (2008) Etude de la diversité génétique et du déséquilibre de liaison au sein de populations améliorées de palmier à huile (*Elaeis guineensis* Jacq.). Thèse de Doctorat, Montpellier SupAgro

Cochard B, Durand-Gasselin T, PalmElit S (2018) Advances in conventional breeding techniques for oil palm. In: Achieving sustainable cultivation of oil palm, vol 1. Burleigh Dodds Science Publishing, pp 133–160

Comstock RE, Robinson HF, Harvey PH (1949) A breeding procedure designed to make maximum use of both general and specific combining ability. *Agron J* 41:360–367

Corley R (2009) How much palm oil do we need? *Environ Sci Pol* 12: 134–139

Corley R, Law I (1997) The future for oil palm clones. In: *Proc Int Planters Conf. Incorp. Soc. Kuala Lumpur*, pp 279–289

Corley R, Tinker P (2016) The oil palm, 5th edn. Wiley-Blackwell, Chichester, p 680

Covarrubias-Pazarán G (2016) Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One* 11:e0156744

Cros D (2014) Etude des facteurs contrôlant l'efficacité de la sélection génomique chez le palmier à huile (*Elaeis guineensis* Jacq.). Montpellier SupAgro, Montpellier 204 p

Cros D, Sánchez L, Cochard B et al (2014) Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population. *Theor Appl Genet* 127:981–994. <https://doi.org/10.1007/s00122-014-2273-3>

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

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

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

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

Cros D, Mbo-Nkoulou L, Bell JM, et al (Under review) Within-family genomic selection in rubber tree increases genetic gain for rubber production. <https://doi.org/10.1016/j.indcrop.2019.111464>

Davidson L (1993) Management for efficient cost-effective and productive oil palm plantations. In: Basiron Y et al (eds) *Proc. 1991 PORIM Int. Oil Palm Conf. Agriculture*. Palm Oil Research Institute of Malaysia, Kuala Lumpur, pp 153–167

- de Carvalho ADF, Fritsche Neto R, Geraldi IO (2008) Estimation and prediction of parameters and breeding values in soybean using REML/BLUP and least squares. *Crop Breed Appl Biotechnol* 8: 219–224
- De Los Campos G, Naya H, Gianola D et al (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182:375–385
- de los Campos G, Hickey JM, Pong-Wong R et al (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193:327–345
- De Souza C (1992) Interpopulation genetic variances and hybrid breeding programs. *Rev Bras Genet* 15:643–643
- Demol J (2002) Amélioration des plantes: application aux principales espèces cultivées en régions tropicales. Presses Agronomiques de Gembloux, Belgique, p 581
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Methodol*: 1–38
- Domonhédou H, Cros D, Nodichao L et al (2018) Enjeux et amélioration de la réduction de l'acidité dans les fruits mûrs du palmier à huile, *Elaeis guineensis* Jacq. (synthèse bibliographique). *Biotechnol Agron Soc Environ* 22:1
- Durán R, Isik F, Zapata-Valenzuela J et al (2017) Genomic predictions of breeding values in a cloned *Eucalyptus globulus* population in Chile. *Tree Genet Genomes* 13:74
- Durand-Gasselin T, Kouame RK, Cochard B et al (2000) Diffusion variétale du palmier à huile (*Elaeis guineensis* Jacq.). *Ol Corps Gras Lipides* 7:207–214
- Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4:250–255
- Falconer D, Mackay T (1996) Introduction to quantitative genetics, 4th edn. Longman, Harlow
- Gallais A (2011) Méthodes de création de variétés en amélioration des plantes. Quae, Versailles 280 p
- Gascon J, De Berchoux C (1964) Caractéristiques de la production de quelques origines d'*Elaeis guineensis* (Jacq.) et de leurs croisements: application à la sélection du palmier à huile. *Oléagineux* 19:75–84
- Gilmour AR, Thompson R, Cullis BR (1995) Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*:1440–1450
- Grattapaglia D (2014) Breeding forest trees by genomic selection: current progress and the way forward. In: *Genomics of plant genetic resources*. Springer, Berlin, pp 651–682
- Habier D, Fernando R, Dekkers J (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177: 2389–2397
- Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12: 186
- Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. *Crop Sci* 49(1):12
- Henderson CR (1950) Estimation of genetic parameters. *International Biometric Soc*, Washington, DC, pp 186–187
- Henderson C (1984) Applications of linear models in animal breeding. Univ Guelph Press Guelph 11:652–653
- Hu X (2015) A comprehensive comparison between ANOVA and BLUP to evaluate location-specific genotype effects for rape cultivar trials with random locations. *Field Crop Res* 179:144–149
- Ithnin M, Xu Y, Marjuni M et al (2017) Multiple locus genome-wide association studies for important economic traits of oil palm. *Tree Genet Genomes* 13:103. <https://doi.org/10.1007/s11295-017-1185-1>
- Jacob F, Cros D, Cochard B, Durand-Gasselin T (2017) Agrigenomics in the breeder's toolbox: latest advances towards an optimal implementation of genomic selection in oil palm. In: *International Seminar on 100 Years of Technological Advancement in Oil Palm Breeding & Seed Production*. ISOPB conference, 13 November 2017, KLCC, Kuala Lumpur, p 21
- Jacquemard JC, Baudoïn L, Noiret JM (1997) Le palmier à huile. In: Charrier A, Jacquot M, Hamon S, Nicolas D (eds) *L'amélioration des plantes tropicales*. CIRAD et ORSTOM, Paris, pp 507–531
- Junaidah J, Rafii M, Chin C, Saleh G (2011) Performance of Tenera oil palm population derived from crosses between Deli Dura and Pisifera from different sources on inland soils. *J Oil Palm Res* 23: 1210–1221
- Kwong QB, Teh CK, Ong AL et al (2016) Development and validation of a high-density SNP genotyping array for African oil palm. *Mol Plant* 9:1132–1141. <https://doi.org/10.1016/j.molp.2016.04.010>
- Kwong QB, Ong AL, Teh CK et al (2017a) Genomic selection in commercial perennial crops: applicability and improvement in oil palm (*Elaeis guineensis* Jacq.). *Sci Rep* 7:2872. <https://doi.org/10.1038/s41598-017-02602-6>
- Kwong QB, Teh CK, Ong AL et al (2017b) Evaluation of methods and marker systems in genomic selection of oil palm (*Elaeis guineensis* Jacq.). *BMC Genet* 18:107
- Lorenz AJ, Chao S, Asoro FG et al (2011) Genomic selection in plant breeding: knowledge and prospects. In: Sparks DL (ed) *Advances in Agronomy*. Academic, Cambridge, pp 77–123
- Marchal A, Legarra A, Tisné S et al (2016) Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests. *Mol Breed* 36:1–13. <https://doi.org/10.1007/s11032-015-0423-1>
- Masani MYA, Izawati AMD, Rasid OA, Parveez GKA (2018) Biotechnology of oil palm: current status of oil palm genetic transformation. *Biocatal Agric Biotechnol* 15:335–347. <https://doi.org/10.1016/j.bcab.2018.07.008>
- Meunier J, Gascon J (1972) Le schéma général d'amélioration du palmier à huile à l'IRHO. *Oléagineux* 27:1–12
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Mrode RA (2005) Linear models for the prediction of animal breeding values, 2nd edn. CABI, Oxfordshire, p 344
- Muñoz F, Sanchez L (2018) breedR: statistical methods for forest genetic resources analysts. <https://github.com/famuvie/breedR>. Accessed Sept 2018
- Muranty H, Jorge V, Bastien C et al (2014) Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. *Tree Genet Genomes* 10:1491–1510
- Noh A, Rafii M, Saleh G et al (2012) Genetic performance and general combining ability of oil palm Deli dura x AVROS pisifera tested on inland soils. *Sci World J* 2012
- Nouy B, Jacquemard J-C, Suryana E, et al (2006) The expected and observed characteristics of several oil palm (*Elaeis guineensis* Jacq.) clones. In: IOPRI (ed). s.n., public, p 17
- Okoye M, Okwuagwu C, Uguru M (2009) Population improvement for fresh fruit bunch yield and yield components in oil palm (*Elaeis guineensis* Jacq.). *Am Eurasian J Sci Res* 4:59–63
- Okwuagwu C, Okoye MN, Okolo E et al (2008) Genetic variability of fresh fruit bunch yield in Deli/dura x tenera breeding populations of oil palm (*Elaeis guineensis* Jacq.) in Nigeria. *J Trop Agric* 46:52–57
- Ong-Abdullah M, Ordway JM, Jiang N et al (2015) Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature* 525:533
- Ooi LC-L, Low E-TL, Abdullah MO et al (2016) Non-tenera contamination and the economic impact of SHELL genetic testing in the Malaysian independent oil palm industry. *Front Plant Sci* 7:771
- Oraguzie NC, Rikkerink EHA, Gardiner SE, de Silva HN (2007) *Association Mapping in Plants*. Springer, Berlin

- Pérez P, de los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. *Plant Genome* 3:106–116
- Piepho H, Möhring J, Melchinger A, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161: 209–228
- Pootakham W, Jomchai N, Ruang-areerate P et al (2015) Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics* 105:288–295
- Potier F, Nouy B, Flori A, et al (2006) Yield potential of oil palm (*Elaeis guineensis* Jacq.) clones: preliminary results observed in the Aek Loba genetic block in Indonesia. *Int. Soc. Oil Palm Breeders Symp. 'Yield potential in oil palm II'*, Phuket, Thailand, 27–28 Nov
- Pszczola M, Strabel T, Mulder H, Calus M (2012) Reliability of direct genomic values for animals with different relationships within and to the reference population. *J Dairy Sci* 95:389–400
- Purba AR, Flori A, Baudouin L, Hamon S (2001) Prediction of oil palm (*Elaeis guineensis*, Jacq.) agronomic performances using the best linear unbiased predictor (BLUP). *Theor Appl Genet* 102:787–792
- R Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna <https://www.R-project.org>. Accessed Sept 2018
- Rafii MY, Isa ZA, Kushairi A et al (2013) Variation in yield components and vegetative traits in Malaysian oil palm (*Elaeis guineensis* Jacq.) dura×pisifera hybrids under various planting densities. *Ind Crop Prod* 46:147–157. <https://doi.org/10.1016/j.indcrop.2012.12.054>
- Rincint R, Laloë D, Nicolas S et al (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics* 192:715–728
- Rincint R, Charcosset A, Moreau L (2017) Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. *Theor Appl Genet*:1–17
- Rival A, Levang P (2014) Palms of controversies: oil palm and development challenges. CIFOR, Jakarta 58 p
- Schnell F, Cockerham C (1992) Multiplicative vs. arbitrary gene action in heterosis. *Genetics* 131:461–469
- Singh R, Low E-TL, Ooi LC-L et al (2013) The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK. *Nature* 500:340
- Soh A (1994) Ranking parents by best linear unbiased prediction (BLUP) breeding values in oil palm. *Euphytica* 76:13–21
- Soh A (1999) Breeding plans and selection methods in oil palm. In: *Symposium on the science of oil palm breeding*. In: Proc. Seminar Science of oil palm breeding. PORIM, Montpellier
- Soh A, Gan H, Wong G et al (2003a) Oil palm genetic improvement. *Plant Breed Rev* 22:165–220
- Soh A, Wong G, Hor T et al (2003b) Estimates of within family genetic variability for clonal selection in oil palm. *Euphytica* 133:147–163
- Soh AC, Wong CK, Ho YW, Choong CW (2010) Oil palm. In: Vollmann J, Rajcan I (eds) *Oil Crops*. Springer New York, New York, pp 333–367
- Soh AC, Mayes S, Roberts JA (2017) *Oil palm breeding: genetics and genomics*. CRC Press, Boca Raton, p 446
- Stuber C, Cockerham CC (1966) Gene effects and variances in hybrid populations. *Genetics* 54:1279
- Ting N-C, Mayes S, Massawe F et al (2018) Putative regulatory candidate genes for QTL linked to fruit traits in oil palm (*Elaeis guineensis* Jacq.). *Euphytica* 214:214. <https://doi.org/10.1007/s10681-018-2296-y>
- Tisné S, Denis M, Cros D et al (2015) Mixed model approach for IBD-based QTL mapping in a complex oil palm pedigree. *BMC Genomics* 16:798
- USDA (2018) Oilseeds: world market and trade. Foreign Agricultural Service, Circular Series November 2018. <https://apps.fas.usda.gov/psdonline/circulars/oilseeds.pdf>. Accessed Nov 2018
- VanRaden PM (2007) Genomic measures of relationship and inbreeding. *Interbull Bull* 37:33–36
- Varshney RK, Roorkiwal M, Sorrells ME (2017) *Genomic selection for crop improvement*, 1st edn. Springer International Publishing, Cham 258 p
- Vitezica ZG, Varona L, Elsen J-M et al (2016) Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. *Genet Sel Evol* 48:6. <https://doi.org/10.1186/s12711-016-0185-1>
- Wang Y, Lin G, Li C, Stothard P (2016) Genotype imputation methods and their effects on genomic predictions in cattle. *Springer Sci Rev* 4:79–98. <https://doi.org/10.1007/s40362-017-0041-x>
- Wei M, Van der Werf JHJ, Brascamp EW (1991) Relationship between purebred and crossbred parameters. *J Anim Breed Genet* 108:262–269. <https://doi.org/10.1111/j.1439-0388.1991.tb00184.x>
- White TL, Hodge GR (1989) *Predicting breeding values with applications in forest tree improvement*. Springer Netherlands, Dordrecht 367 p
- Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS (2017) Genomic selection in dairy cattle: the USDA experience. *Annu Rev Anim Biosci* 5:309–327. <https://doi.org/10.1146/annurev-animal-021815-111422>
- Wong C, Bernardo R (2008) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor Appl Genet* 116:815–824
- Xavier A, Muir WM, Craig B, Rainey KM (2016) Walking through the statistical black boxes of plant breeding. *Theor Appl Genet* 129: 1933–1949

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



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

Achille Nyouma^{a,b}, Joseph Martin Bell^a, Florence Jacob^c, Virginie Riou^{d,e}, Aurore Manez^{d,e}, Virginie Pomiès^{d,e}, Leifi Nodichao^{e,f}, Indra Syahputra^g, Dadang Affandi^g, Benoit Cochard^c, Tristan Durand-Gasselín^c, David Cros^{b,d,e,*}

^a Department of Plant Biology, Faculty of Science, University of Yaoundé 1, Yaoundé, Cameroon

^b CETIC (African Center of Excellence in Information and Communication Technologies), University of Yaoundé 1, Yaoundé, Cameroon

^c PalmElit SAS, 34980, Montferrier sur Lez, France

^d CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP, F-34398, Montpellier, France

^e AGAP, University of Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France

^f INRAB, CRA-PP, Pobè, Benin

^g P.T. SOCFINDO Medan, Medan, Indonesia

ARTICLE INFO

Keywords:

Elaeis guineensis

Genomic selection

Ortets

Clonal selection

Genotyping-by-sequencing

Prediction accuracy

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 0.08 to 0.70 for ortet candidates without data records, depending on trait, SNP dataset and modeling. ASGM was better (on average slightly more accurate, less sensitive to SNP dataset and simpler), although PSAM appeared interesting for a few traits. With ASGM, the number of SNPs had to reach 7,000, while the percentage of missing data per SNP was of secondary importance, and 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.

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

* Corresponding author at: CETIC (African Center of Excellence in Information and Communication Technologies), University of Yaoundé 1, Yaoundé, Cameroon.
E-mail address: david.cros@cirad.fr (D. Cros).

<https://doi.org/10.1016/j.plantsci.2020.110547>

Received 28 January 2020; Received in revised form 14 April 2020; Accepted 1 June 2020

Available online 03 June 2020

0168-9452/ © 2020 Elsevier B.V. All rights reserved.

[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 micro-propagation 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. Nouy et al. [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 pre-selection 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

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

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.

2.2. Phenotyping

All the individuals, i.e. the training hybrid crosses, the 42 hybrid ortets and their ramets, were phenotyped for eight traits. Five traits were assessed for bunch quality: 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); and three traits for bunch production: bunch number (BN), average bunch weight (ABW), and total bunch production (FFB). For quality traits, data were collected when plants were from five to nine years old at ALT and from six to nine years old at AK1. For production traits, data were collected when the plants were from three to seven years old in both sites.

2.3. Genotyping

Molecular data were obtained by GBS [25,26] for the 42 ortets, 93 Deli and 91 La Mé parents of the training hybrid crosses (Table 1). Ortets genotypes were obtained from two or three samples collected on different ramets (thus allowing controlling the legitimacy of the ramets). DNA extraction and GBS were performed as described in Cros et al. [15], using the *Pst*I and *Hha*I restriction enzymes. The raw fastq sequence data were processed with Tassel GBS v. 5.2.44 [27], using the Bowtie2 software for alignment [28], and VCFtools 0.1.14 [29]. The indels were discarded, the datapoints with depth below five were set to missing, the SNPs that were not biallelic, with more than 75 % of missing data or on the unassembled part of the genome were discarded (see Cros et al. [15] for more details about SNP calling and filtering). This resulted in a dense genome covering with 15,054 SNPs. The average percentage of missing data was 23.08 % (3.64 %–43.42 % per individual). To explain the 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, for the missing parents that were obtained through selfing, the selfed grandparents were used in the pedigree instead of the actual parents, as grandparental DNA was available (for the other steps of the analysis that required a pedigree, the real pedigree was used). As some ortets remained with one parent that was not genotyped and that did not originate from a selfing, we used a home-made R script to recover the parental origin of ortet phases. For each ortet, this script considered the two phases, one after another, and checked all along the genome if similar blocks of consecutive SNPs were found in the Deli and La Mé parent. Each ortet phase was finally assigned to the parental population with the highest number of SNP blocks specific to the population that were found on the considered ortet phase..

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.

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

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

In all cases, the models were trained with the phenotypic data of ALT hybrids and the genomic data of their parents, and the genetic values of the 42 validation clones were predicted. For all the models mentioned above, no phenotypic data of the validation clones were

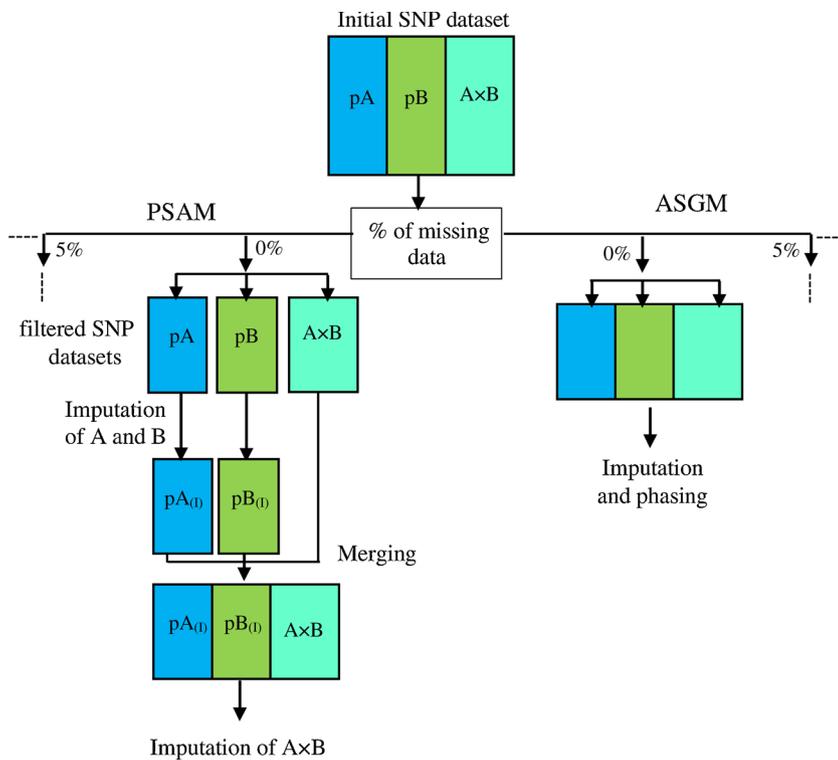


Fig. 1. Imputation and phasing scheme for the production of the SNP datasets used for genomic predictions with the two models PSAM (population-specific effects of SNP alleles model) and ASGM (across-population SNP genotype model). pA, pB, A × B: Deli parents, La Mé parents and Deli × La Mé hybrid ortets,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

provided to the prediction models. This corresponds to a breeding situation where predictions are made for immature individuals (e.g. nursery plantlets belonging to crosses that were not evaluated in progeny-tests but were produced by mating the best parents selected at the end of the progeny-tests). However, ortet selection can also be made within the crosses evaluated in progeny tests. In this case, the ortet candidates have phenotypic data records, which should be taken into consideration along with their SNP data when predicting their clonal value. This was evaluated with the G_ASGM_A model, simply including the adjusted phenotypic value of the validation ortets (see below) to the phenotypic dataset used to train 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:

$$y = X\beta + Z_1g_i + Z_2g_{Deli \times LM} + Z_3b + Z_4p + \varepsilon$$

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), $g_i \sim N(0, H_i\sigma_{ai}^2)$ the individual additive genetic effects, $g_{Deli \times LM} \sim N(0, H_{Deli \times LM}\sigma_{dDeli \times LM}^2)$ the genetic dominance effects, $b \sim N(0, I\sigma_b^2)$ the incomplete block effect, and $p \sim N(0, I\sigma_p^2)$ the elementary plot effects. X , Z_1 , Z_2 , Z_3 and Z_4 are the incidence matrices associated to β , g_i , $g_{Deli \times LM}$, b and p respectively. $H_i\sigma_{ai}^2$

and $H_{Deli \times LM}\sigma_{dDeli \times LM}^2$ are the variance-covariance matrices associated with g_i and $g_{Deli \times LM}$, respectively. σ_{ai}^2 and $\sigma_{dDeli \times LM}^2$ are the additive and dominance variances, respectively. $\varepsilon \sim N(0, I\sigma_\varepsilon^2)$ is the vector of residual effects and 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 H_i 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 H_i matrices subsequently in this paper will refer to matrices combining genealogical and genomic information. H_i^{-1} is the inverse of H_i , computed according to Misztal et al. [33] as:

$$H_i^{-1} = A_i^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G_i^{-1} - A_{i22}^{-1} \end{bmatrix}$$

where G_i^{-1} and A_{i22}^{-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 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

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

(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 G was obtained as: $G = \frac{XX'}{2 \sum_{l=1}^{m \text{ SNPs}} p_l(1-p_l)}$, with $X = Z - P$, X' the transpose of matrix X , Z the SNP matrix containing the number of copies of the minor allele at an SNP (ranging from 0 to 2), P a matrix given by $P = 2p_l$, and p_l the frequency of the minor allele at SNP l [34]. $H_{Deli \times LM}$ 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. $H_{Deli \times LM}^{-1}$ was computed following the same method as H_i^{-1} except that the additive relationship matrices were replaced by the dominance relationship matrices. The realized dominance relationship matrix G_D was computed according to Su et al. [35] as: $G_D = \frac{II'}{2 \sum_{l=1}^m p_l q_l (1 - 2p_l q_l)}$, with II' the $n \times m$ matrix (n : number of hybrid crosses and clones and m : number of SNPs) of heterozygosity coefficients with element $II_{kl} = 0 - p_l q_l$ if clone or ortet k is homozygous and $II_{kl} = 1 - p_l q_l$ 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, H_i was replaced by the additive genealogical relationship matrix A_i and, for P_ASGM_AD, $H_{Deli \times LM}$ 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:

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

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 of the hybrid crosses and the ortets from the Deli and La Mé populations, respectively, and $g_{Deli \times LM} \sim N(0, H_{Deli \times LM} \sigma_{g_{Deli \times LM}}^2)$ the dominance effects of the crosses and clones. X , Z_1 , Z_2 , Z_3 , Z_4 , Z_5 are the incidence matrices associated to β , g_{Deli} , g_{LM} , $g_{Deli \times LM}$, b and p , respectively. $H_{Deli} \sigma_{g_{Deli}}^2$, $H_{LM} \sigma_{g_{LM}}^2$ and $H_{Deli \times LM} \sigma_{g_{Deli \times LM}}^2$ are the variance-covariance matrices associated to 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 and La Mé populations, respectively, and $\sigma_{g_{Deli \times LM}}^2$ is the genetic dominance variance of crosses and clones. 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. H_{LM} is defined similarly for the La Mé population. To build H_{Deli} , we created first the matrix of additive realized relationships of Deli parents G_{Deli} (incorporating the Deli parents of the training and validation hybrid crosses and clones) as follows [49]: $G_{Deli} = \begin{bmatrix} G_{Deli, Deli}^{Deli, Deli} & G_{Deli, Deli \times LM}^{Deli, Deli \times LM} \\ G_{Deli, Deli \times LM, Deli}^{Deli \times LM, Deli} & G_{Deli, Deli \times LM, Deli \times LM}^{Deli \times LM, Deli \times LM} \end{bmatrix}$

with, $G_{Deli, Deli}^{Deli, Deli} = (Z_{Deli} - 2p_{Deli}1')(Z_{Deli} - 2p_{Deli}1)'$, $G_{Deli, Deli \times LM}^{Deli, Deli \times LM} = (Z_{Deli} - 2p_{Deli}1')(Z_{Deli \times LM} - p_{Deli}1)'$ and $G_{Deli, Deli \times LM, Deli \times LM}^{Deli \times LM, Deli \times LM} = (Z_{Deli \times LM} - p_{Deli}1')(Z_{Deli \times LM} - p_{Deli}1)'$. Z_{Deli} and $Z_{Deli \times LM}$ are the matrices containing the number of copies of reference allele in the genotyped Deli parents (coded as 0, 1 or 2) and in the Deli haplotype of clones (coded as 0 or 1), respectively, p_{Deli} is the vector containing the allele frequencies based on SNP genotypes of Deli parents and Deli haplotype in clones and 1 is a vector of ones. G_{Deli} was then adjusted to be in the same scale and compatible with the genealogical additive relationship matrix of the clones and the genotyped Deli parents A_{Deli22} , according to Christensen et al. [50] and Xiang et al. [49], and using weight 0.001, to give the $G_{Deli_{iw}}$ matrix. Then the inverse of H_{Deli} was constructed as:

$H_{Deli}^{-1} = A_{Deli}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G_{Deli_{iw}}^{-1} - A_{Deli22}^{-1} \end{bmatrix}$, with A_{Deli}^{-1} the inverse of the genealogical relationship matrix of all the Deli parents and clones. H_{LM} was created following the same procedure as H_{Deli} . $H_{Deli \times LM}$ 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: $H_{Deli \times LM} = H_{Deli} \otimes H_{LM}$, with \otimes the Kronecker product.

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

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

2.7. Prediction accuracies

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

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

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

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

2.9. Accuracy of phenotypic selection before clonal trials

To evaluate the possibility of using GS instead of the current phenotypic selection (PS) to select the hybrid individuals to test in the clonal trials, the PS accuracy was computed for each trait. It 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

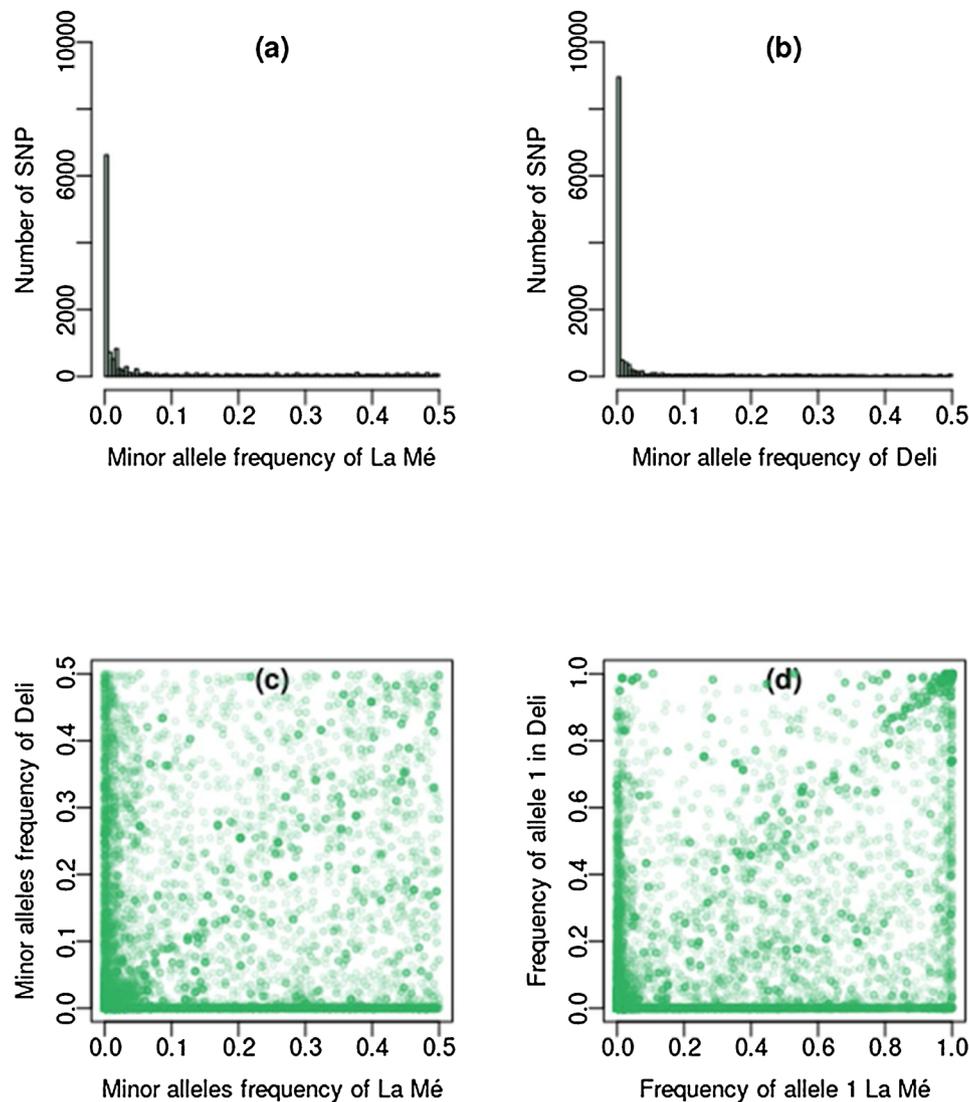


Fig. 2. Distribution of minor allele frequency (MAF) in La Mé (a) and Deli (b) populations, and correlation of MAF (c) and frequency of alternate alleles between La Mé and Deli (d). In (c) and (d) panels, each dot represents an SNP.

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 0.08 to 0.70 depending on prediction model, trait and SNP dataset (Fig. 3) for additive models (G_ASJM_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,

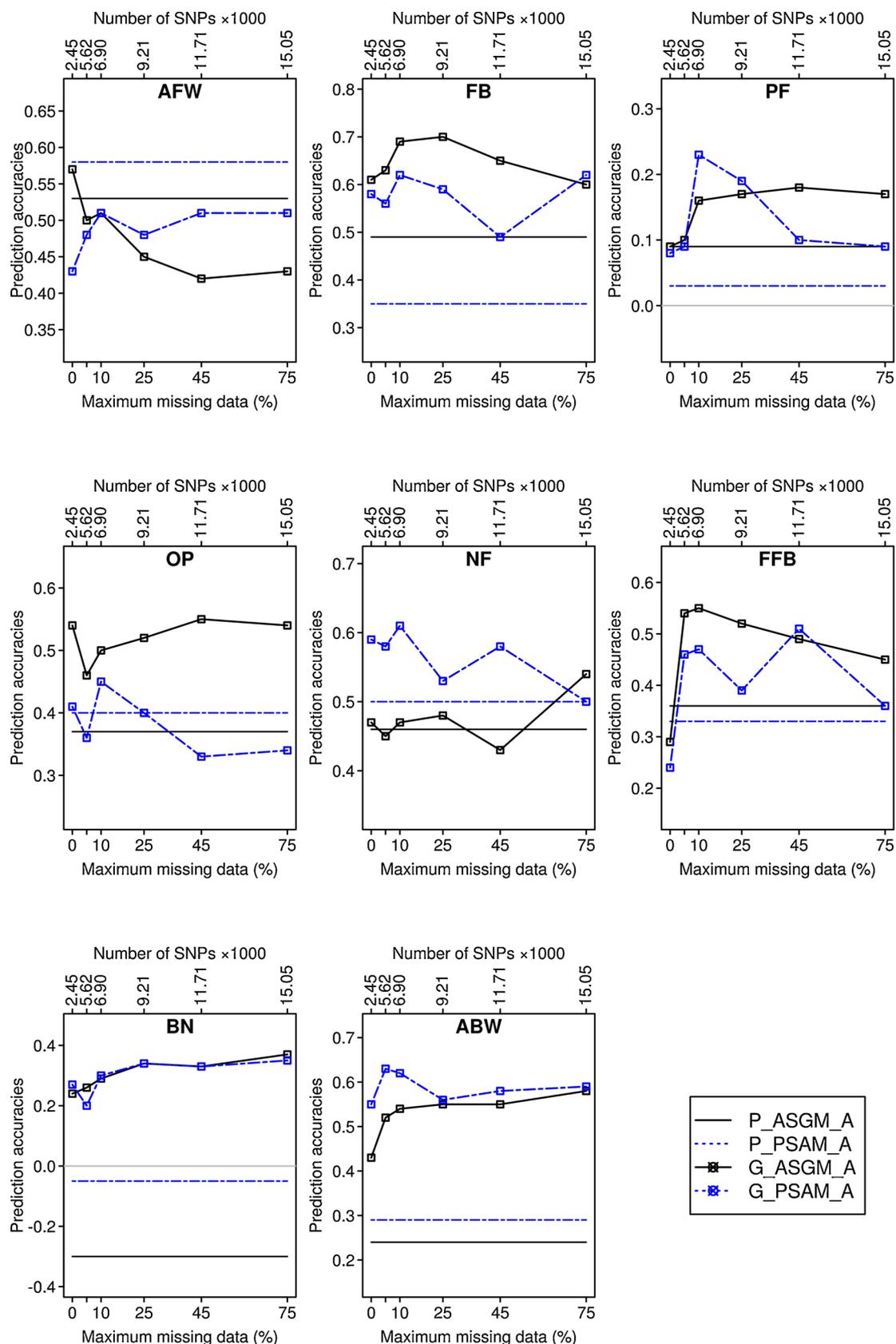


Fig. 3. Prediction accuracies according to traits, SNP datasets and prediction models.

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 (0.43), 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 0.13 (PF) to 0.59 (ABW) for G_PSAM_A. G_ASGM_A obtained a mean prediction accuracy greater than

Table 3
Mean prediction accuracies according to trait and prediction model.

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	0.49	0.57 (0 %)	0.51 (10 %/45 %/75 %)
FB	0.65	0.58	0.70 (25 %)	0.62 (10 %/75 %)
PF	0.14	0.13	0.18 (45 %)	0.23 (10 %)
OP	0.52	0.38	0.55 (45 %)	0.45 (10 %)
NF	0.47	0.57	0.54 (75 %)	0.61 (10 %)
FFB	0.47	0.41	0.55 (10 %)	0.51 (45 %)
BN	0.31	0.30	0.37 (75 %)	0.35 (75 %)
ABW	0.53	0.59	0.58 (75 %)	0.63 (5 %)
Mean	0.45	0.43	0.51	0.49

Bunch production: bunch number (BN), average bunch weight (ABW) and total bunch production (FFB); bunch quality: 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); genomic prediction models: across-population SNP genotype models (ASGM_A), population-specific effects of SNP alleles models (PSAM_A). Values in brackets indicate the corresponding SNP dataset, defined on its maximum percentage of missing data.

G_PSAM_A for five traits out of eight, with G_PSAM_A being on average slightly more accurate than G_ASGM_A for AFW, NF and ABW (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 0.23 (PF) to 0.63 (ABW) for G_PSAM_A (Table 3), and G_ASGM_A was again more often better than G_PSAM_A (with G_PSAM_A being more accurate for PF, NF and ABW). Considering the different SNP datasets and traits, G_ASGM_A gave higher prediction accuracy than G_PSAM_A in 58.3% of the cases, with the largest differences in prediction accuracy in favor of G_ASGM_A, up to 0.22 with OP at $p_{max} = 45\% \cdot n_{SNP} = 11,707$ (although they were non-significant) (Fig. 3 and Table 4). Significant differences were only found in favor of G_PSAM_A, but they were scarce (i.e. only for NF in three SNP datasets, $p_{max} = 5\% \cdot n_{SNP} = 5,620$, $p_{max} = 10\% \cdot n_{SNP} = 6,898$ and $p_{max} = 45\% \cdot n_{SNP} = 11,707$). Despite the overall lower prediction accuracies of G_PSAM_A compared to G_ASGM_A, G_PSAM_A was the most accurate method for ABW and NF with all the SNP datasets, except for NF with $p_{max} = 75\% \cdot n_{SNP} = 15,054$. G_ASGM_A, therefore, appeared to be a the best approach (i.e. generally more accurate, in addition to being easier to implement) for predicting clonal values for oil palm yield components, although G_PSAM_A could be worthwhile for some traits (ABW and NF here).

Prediction accuracies could be broadly improved when relationship matrices were computed using SNPs (G_ASGM_A and G_PSAM_A) instead of genealogical data (control pedigree-based models P_ASGM_A and P_PSAM_A), in particular for three traits FB, BN and ABW. The maximum prediction accuracies of GS over all SNP datasets outperformed pedigree-based models for seven traits out of eight (except

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 $p_{max}\% \cdot n_{SNP}$. 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 %-2447	$G_{ASGM_A} - G_{PSAM_A}$	0.14	0.03	0.01	0.13	-0.12	0.05	-0.03	-0.12
5 %-5620	$G_{ASGM_A} - G_{PSAM_A}$	0.02	0.07	0.01	0.10	-0.13*	0.08	0.06	-0.11
10 %-6898	$G_{ASGM_A} - G_{PSAM_A}$	0.00	0.07	-0.07	0.05	-0.14*	0.08	-0.01	-0.08
25 %-9,205	$G_{ASGM_A} - G_{PSAM_A}$	-0.03	0.11	-0.02	0.12	-0.05	0.13	0.00	-0.01
45 %-11,707	$G_{ASGM_A} - G_{PSAM_A}$	-0.09	0.16	0.08	0.22	-0.15*	-0.02	0.00	-0.03
75 %-15,054	$G_{ASGM_A} - G_{PSAM_A}$	-0.08	-0.02	0.08	0.20	0.04	0.09	0.02	-0.01

for AFW with G_PSAM_A) (Table 5 and Fig. 3). The largest difference was observed in BN for $p_{max} = 75\% \cdot n_{SNP} = 15,054$, with G_ASGM_A accuracy being 0.67 higher than P_ASGM_A. Significant differences between GS models and their pedigree-based control models were found for five traits, with four traits (FB, OP, BN and ABW) where GS was the best and one trait (AFW) where pedigree-based models were more accurate (Table 5). The percentage of combinations of SNP datasets and traits where G_ASGM_A was more accurate than its control pedigree-based version reached 83.3%, against only 64.6% for G_PSAM_A.

The SNP dataset affected the prediction accuracy differently according to the trait and the model. 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 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 two traits (AFW and BN). For the other traits, prediction accuracies remained stable or tended to decrease with increasing marker density and maximum percentage of missing SNP data.

However, the use of a different SNP dataset for each combination of trait and model seems unrealistic for the practical application of GS. Therefore, in order to identify the optimal SNP dataset(s) that would maximize GS accuracy, we computed for each GS prediction model and SNP dataset the mean prediction accuracy over the traits. For G_ASGM_A, this value increased with the SNP density (0.41 with SNP dataset $p_{max} = 0\% \cdot n_{SNP} = 2,447$ and 0.43 with $p_{max} = 5\% \cdot n_{SNP} = 5,620$), before plateauing at 0.46 with the subsequent SNP datasets. This shows that, for G_ASGM_A, the number of SNPs was of greater importance than the percentage of missing data per SNP. Mean prediction accuracy over the SNP datasets forming the plateau ranged from 0.17 (PF) to 0.66 (FB), and were close to the highest accuracies achieved over all the SNP datasets (Table 3). For G_ASGM_A, there was therefore a minimum of 6,898 SNPs required to reach maximum prediction accuracy on average over all traits. For G_PSAM_A, the results differed, with a peak in mean prediction accuracy at 0.47 with SNP dataset $p_{max} = 10\% \cdot n_{SNP} = 6,898$ and mean prediction accuracy decreasing when less SNPs were used, falling to 0.39 with $p_{max} = 0\% \cdot n_{SNP} = 2,447$, and decreasing when there were more missing data, falling to 0.41 with $p_{max} = 75\% \cdot n_{SNP} = 15,054$. This shows that G_PSAM_A was more sensitive to the SNP dataset than G_ASGM_A, making again G_PSAM_A less appealing. Therefore, for the final part of the study, we decided to focus on G_ASGM_A.

3.3. Comparison of prediction accuracies of PS and GS

Fig. 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 +

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	<i>P</i> _ASGM_A – <i>G</i> _ASGM_A	-0.04	-0.12	0.00	-0.17	-0.01	0.07	-0.53**	-0.19
	<i>P</i> _PSAM_A – <i>G</i> _PSAM_A	0.15	-0.23*	-0.05	-0.01	-0.09	0.09	-0.32*	-0.26
5 %-5,620	<i>P</i> _ASGM_A – <i>G</i> _ASGM_A	0.03	-0.14	-0.01	-0.09	-0.01	-0.18	-0.56**	-0.28*
	<i>P</i> _PSAM_A – <i>G</i> _PSAM_A	0.10	-0.21	-0.06	-0.04	-0.08	-0.13	-0.25	-0.34*
10 %-6,898	<i>P</i> _ASGM_A – <i>G</i> _ASGM_A	0.02	-0.20*	-0.07	-0.13	-0.01	-0.18	-0.59**	-0.30*
	<i>P</i> _PSAM_A – <i>G</i> _PSAM_A	0.07	-0.27*	-0.20	-0.05	-0.11	-0.14	-0.35*	-0.33*
25 %-9,059	<i>P</i> _ASGM_A – <i>G</i> _ASGM_A	0.08	-0.20*	-0.08	-0.15	-0.02	-0.16	-0.64***	-0.30**
	<i>P</i> _PSAM_A – <i>G</i> _PSAM_A	0.10	-0.24*	-0.16	0.00	-0.03	-0.06	-0.39**	-0.27*
45 %-11,425	<i>P</i> _ASGM_A – <i>G</i> _ASGM_A	0.11	-0.15	-0.09	-0.18*	0.03	-0.13	-0.62***	-0.30**
	<i>P</i> _PSAM_A – <i>G</i> _PSAM_A	0.07	-0.14	-0.07	0.07	-0.08	-0.18	-0.38*	-0.29*
75 %-15,054	<i>P</i> _ASGM_A – <i>G</i> _ASGM_A	0.10*	-0.11	-0.08	-0.17	-0.08	-0.09	-0.67***	-0.34***
	<i>P</i> _PSAM_A – <i>G</i> _PSAM_A	0.07	-0.27**	-0.06	0.06	0.00	-0.03	-0.40*	-0.30*

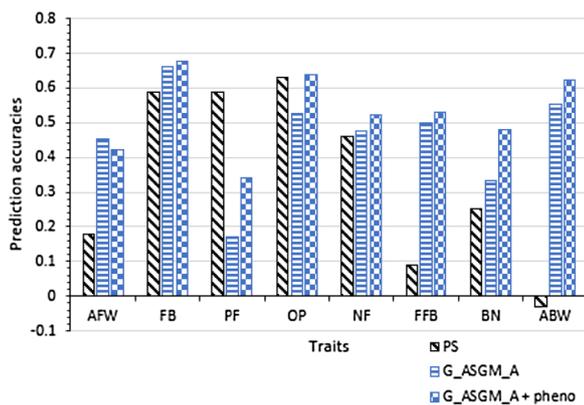


Fig. 4. Prediction accuracies of phenotypic selection (PS) and of the *G*_ASGM_A model without phenotypic data (*G*_ASGM_A) and with phenotypic data (*G*_ASGM_A + pheno) of ortets, on average over the best SNP datasets, and according to trait.

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

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*_ASGM_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 should 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 pre-selected. 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 10–50 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 palm 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 usually did not perform better than G_ASGM_A, except for ABW and NF. We hypothesized that this was the consequence of stronger differences among Deli and La Mé populations at the QTLs controlling ABW and NF than 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, and noting that, although G_PSAM_A was not better than G_ASGM_A for BN, their results were actually very similar for this trait. 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. The possibility for G_PSAM_A to outperform G_ASGM_A 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 G_PSAM_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.

Ibáñez-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 300–400 generations ago, and a large training population with 2,000–8,000 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). In our study, the only SNP dataset where G_PSAM_A outperformed G_ASGM_A on average over all traits was a dataset with intermediate number of SNPs and intermediate percentage of missing data per SNP, $p_{max} = 10\% - nSNP = 6,898$, with mean G_PSAM_A prediction accuracy of 0.47 against 0.46 for G_ASGM_A. This result therefore differs from those of Technow et al. [22] and Lopes et al. [45], likely as a consequence of the fact that, in our study, SNP density varied with SNP quality, with higher SNP numbers meaning a higher percentage of missing data. This indicates that the SNP dataset must be chosen carefully before applying G_PSAM_A. From this point of view, G_ASGM_A appeared advantageous, as its mean accuracy over the traits remained at its maximum once sufficient SNP density was reached, regardless of the percentage of missing data. The fact that for G_ASGM_A the number of SNPs was of

greater importance than the percentage of missing data per SNP indicates that Beagle 4.0 efficiently imputed the missing data. Therefore, the existence of an optimal SNP dataset for G_PSAM_A suggests that phasing errors increase with the percentage of missing data per SNP and when decreasing the marker density.

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 7,000–15,000 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 one trait (AFW), while that never happened with G_ASGM_A. Also, the overall inferiority of G_PSAM_A to G_ASGM_A occurred while P_PSAM_A was actually better than P_ASGM_A for five traits out of eight. This looks contradictory and suggests that the performance of G_PSAM_A could have been reduced by phasing errors as aforementioned. Also, many studies comparing G_ASGM_A and G_PSAM_A were carried out by simulation with known phases [22,42,43], and therefore possible phasing errors in our study could also be the cause of the discrepancies observed between our results and the results obtained in simulation studies. Investigating other phasing approaches seems therefore of interest in the oil palm context.

4.3. Genotyped individuals for training

In this study, to make GS predictions more cost-effective, the genotypes of the phenotyped hybrid individuals constituting the training set were reconstructed using the molecular data of their parents, with G_ASGM, or not used in the model, with G_PSAM. Both modeling approaches therefore assume that the mean genotype in a hybrid family (i.e. the mean number of copies of the minor allele over the individuals making the family) expected from the parental genotypes is the same as the actual mean genotype. Nevertheless, in the case of allele segregation distortion at a locus, the mean genotype in a hybrid family would significantly deviate from the mean genotype expected from the parental genotypes, and this could reduce the GS accuracy. Indeed, high numbers of distorted markers can be found in plants: Zuo et al. [46] and Li et al. [47] found more than 10 % of markers (SNP and SSR) significantly distorted. For future studies, it would be of great interest to compare the approach used here with predictions made using real hybrid genotypes, and to measure the differences in terms of GS accuracy and cost.

4.4. Prediction of dominance effects

GS prediction accuracies were not significantly enhanced by adding dominance effects. Including dominance effects in the statistical model sometimes slightly increased or reduced accuracies, depending on the traits and the SNP datasets, revealing a negligible genetic dominance variance captured by the model compared to the total genetic variance, as already observed with genomic predictions for performances of oil

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

5. Conclusion

This work showed that GS can largely improve clonal selection in oil palm (*Elaeis guineensis*). GS prediction accuracies for ortets without phenotypic data records extended from 0.08 to 0.7 according to the trait, GS model and SNP dataset. The G_ASGM_A approach was better for predicting clonal values than G_PSAM_A, as it was on average slightly more accurate, less sensitive to SNP dataset (i.e. SNP density and percentage of missing data) and easier to implement. However, G_PSAM_A appeared interesting for ABW and NF traits. The G_ASGM_A model required at least 7,000 SNPs to perform best, with the percentage of missing data per SNP being of secondary importance. In these conditions, G_ASGM_A gave higher prediction accuracies than current phenotypic selection for six traits out of eight.

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

Data availability

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

Author contributions

AN carried out data analysis, under the supervision of DC and JMB. The paper was written by AN and DC, with the help of FJ and JMB. IS, DA and LN supervised the production of the plant material, the field trials and the collection of the phenotypic data, with help of BC and TDG. BC, TDG, IS and DA designed the field experiments. The molecular data were generated by AM, VR and VP.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge SOCFINDO (Indonesia), CRAPP (Benin) and PalmElit (France) for 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.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:<https://doi.org/10.1016/j.plantsci.2020.110547>.

References

- [1] USDA, <http://www.fas.usda.gov/data/oilseeds-world-markets-and-trade>. Accessed 13 January 2020, 2020.
- [2] R.H.V. Corley, P.B. Tinker, *The Oil Palm*, 5th ed., Wiley-Blackwell, Chichester, UK, 2016, <https://doi.org/10.1002/9781118953297>.
- [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 (Accessed 23 October 2014), http://www.cifor.org/publications/pdf_files/Books/BLevang1401.pdf.
- [6] R. Corley, I. Law, *The Future of Oil Palm Clones*, (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 Rep.* 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-Gasselini, The Expected and Observed Characteristics of Several Oil Palm (*Elaeis guineensis* Jacq.) Clones, IOPRI, International Oil Palm Conference (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, *Front. Plant Sci.* 9 (2018) 1693.
- [14] C. Wong, R. Bernardo, Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations, *Theor. Appl. Genet.* 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 Genet. 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.), *Sci. Rep.* 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 Genet. 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, *Ind. Crops Prod.* 138 (2019) 111464, <https://doi.org/10.1016/j.indcrop.2019.111464>.
- [20] N. Ibáñez-Escriche, R. Fernando, A. Toosi, J. Dekkers, Genomic selection of purebreds for crossbred performance, *Genet. Sel. Evol.* 41 (2009) 12.
- [21] C. Stuber, C.C. Cockerham, Gene effects and variances in hybrid populations, *Genetics* 54 (1966) 1279.
- [22] F. Technow, C. Riedelheimer, Tobias A. Schrag, Albrecht E. 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>.
- [23] R.H.V. Corley, P.B. Tinker, *The Oil Palm*, 5th ed., Wiley-Blackwell, Chichester, UK, 2016, <https://doi.org/10.1002/9781118953297>.
- [24] F. Potier, B. Nouy, A. Flori, J. Jacquemard, H. Edyana Suryana, T. Durand-Gasselini, Yield Potential of Oil Palm (*Elaeis guineensis* Jacq) Clones: Preliminary Results Observed in the Aek Loba Genetic Block in Indonesia, IOPRI, International Oil Palm Conference (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, *Front. Plant Sci.* 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, *Nat. 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, *Am. J. Hum. Genet.* 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, *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 Bull.* (2009) 240.
- [34] P.M. VanRaden, Efficient methods to compute genomic predictions, *J. Dairy Sci.* 91 (2008) 4414–4423.
- [35] G. Su, O.F. Christensen, T. Ostensen, 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, *Psychol. Bull.* 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, *Mol. Breed.* 38 (2018) 89, <https://doi.org/10.1007/s11032-018-0850-x>.
- [39] R.H.V. Corley, P.B. Tinker, *Vegetative propagation and biotechnology, 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-Gasselini, 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, *Genet. Sel. Evol.* 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, *Genet. Sel. Evol.* 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, *Genet. Sel. Evol.* 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* 123 (2019) 579–592, <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 Res. Int.* 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.
- [49] T. Xiang, B. Nielsen, G. Su, A. Legarra, O.F. Christensen, Application of single-step genomic evaluation for crossbred performance in pig, *J. Anim. Sci.* 94 (2016) 936–948.
- [50] O.F. Christensen, P. Madsen, B. Nielsen, T. Ostensen, G. Su, Single-step methods for genomic evaluation in pigs, *Animal* 6 (10) (2012) 1565–1571, <https://doi.org/10.1017/S1751731112000742>.