

THE UNIVERSITY OF YAOUNDE I

FACULTY OF SCIENCE

POSTGRADUATE SCHOOL IN LIFE
SCIENCE, HEATH & ENVIRONMENTAL
SCIENCES



UNIVERSITÉ DE YAOUNDÉ I

FACULTÉ DES SCIENCES

CENTRE DE RECHERCHE ET DE
FORMATION DOCTORALE EN
SCIENCES DE LA VIE, SANTE ET
ENVIRONNEMENT

DEPARTMENT OF PLANT BIOLOGY
DÉPARTEMENT DE BIOLOGIE ET PHYSIOLOGIE VÉGÉTALES

**Characterization of genome properties in oil
palm (*Elaeis guineensis* Jacq.) breeding
populations for better palm oil yield**

Thesis

Submitted in partial fulfillment for the requirements for award of
Doctor of Philosophy Degree (Ph.D) in Plant Biology
Option: Plant Biotechnology

By

ESSUBALEW Getachew Seyum

MSc. in Horticulture

Registration Number: 18W4586



Defended on June 16, 2023 in front of the Jury members as follows:

- Président** : AMBANG Zachée, Pr., Université de Yaoundé I
Rapporteurs : BELL Joseph Martin, Pr., Université de Yaoundé I
NGALLE Hermine BILLE, M.C., Université de Yaoundé I
CROS David, PhD/HDR, CIRAD - Montpellier
Membres : NDONGO BEKOLO, M.C., Université de Yaoundé I
NGONKEU M. Eddy L., M.C., Université de Yaoundé I
KOUAM Eric Bertrand, M.C., Université de Dschang

Year 2023



DEPARTEMENT DE BIOLOGIE ET PHYSIOLOGIE VEGETALES
DEPARTMENT OF PLANT BIOLOGY

Yaoundé, le 20 JUIN 2023

ATTESTATION DE CORRECTION

Nous soussignés, membres du Jury de soutenance de la thèse de **Doctorat/PhD en Biologie des Organismes Végétaux**, Option: **Biotechnologies Végétales**, Spécialité : **Génétique et Amélioration des Plantes**, soutenue le **16 juin 2023** par Monsieur **SEYUM Esubalew Getachew**, MSc en Horticulture, Matricule : **18W4586**, thèse intitulée « **Characterization of genome properties in oil palm (*Elaeis guineensis* Jacq.) breeding populations for better palm oil yield** », certifions qu'il a effectué les corrections conformément aux remarques et recommandations du Jury.

En foi de quoi, nous lui délivrons cette attestation de correction pour servir et valoir ce que de droit./-

BELL Joseph Martin,
Professeur

Rapporteurs de la thèse

NGALLE Hermine BILLE,
Maître de Conférences

CROS David,
PhD/HDR

NDONGO BEKOLO,
Maître de Conférences

Membres du Jury


NGONKEU M. Eddy L.,
Maître de Conférences

KOUAM Eric Bertrand,
Maître de Conférences

Président du Jury

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 2021/2022

(Par Département et par Grade)

DATE D'ACTUALISATION 22 juin 2022

ADMINISTRATION

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

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

VICE-DOYEN / DSSE : NYEGUE Maximilienne Ascension, *Professeur*

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

Chef Division Administrative et Financière : NDOYE FOE Florentine Marie Chantal, *Maître de Conférences*

Chef Division des Affaires Académiques, de la Recherche et de la Scolarité DAARS : AJEAGAH Gideon AGHAINDUM, *Professeur*

1- DÉPARTEMENT DE BIOCHIMIE (BC) (39)			
N°	NOMS ET PRÉNOMS	GRADE	OBSERVATIONS
1.	BIGOGA DAIGA Jude	Professeur	En poste
2.	BOUDJEKO Thaddée	Professeur	En poste
3.	FEKAM BOYOM Fabrice	Professeur	En poste
4.	FOKOU Elie	Professeur	En poste
5.	KANSCI Germain	Professeur	En poste
6.	MBACHAM FON Wilfred	Professeur	En poste
7.	MOUNDIPA FEWOU Paul	Professeur	Chef de Département
8.	OBEN Julius ENYONG	Professeur	En poste
9.	ACHU Merci BIH	Maître de Conférences	En poste
10.	ATOGHO Barbara MMA	Maître de Conférences	En poste
11.	AZANTSA KINGUE GABIN BORIS	Maître de Conférences	En poste
12.	BELINGA née NDOYE FOE F. M. C.	Maître de Conférences	<i>Chef DAF / FS / UYI</i>
13.	DJUIDJE NGOUNOUE Marceline	Maître de Conférences	En poste
14.	EFFA ONOMO Pierre	Maître de Conférences	En poste

15.	EWANE Cécile Annie	Maître de Conférences	En poste
16.	KOTUE TAPTUE Charles	Maître de Conférences	En poste
17.	MOFOR née TEUGWA Clotilde	Maître de Conférences	<i>Doyen FS / UDs</i>
18.	NANA Louise épouse WAKAM	Maître de Conférences	En poste
19.	NGONDI Judith Laure	Maître de Conférences	En poste
20.	NGUEFACK Julienne	Maître de Conférences	En poste
21.	NJAYOU Frédéric Nico	Maître de Conférences	En poste
22.	TCHANA KOUATCHOUA Angèle	Maître de Conférences	En poste
23.	AKINDEH MBUH NJI	Chargé de Cours	En poste
24.	BEBEE Fadimatou	Chargée de Cours	En poste
25.	BEBOY EDJENGUELE Sara Nathalie	Chargé de Cours	En poste
26.	DAKOLE DABOY Charles	Chargé de Cours	En poste
27.	DJUIKWO NKONGA Ruth Viviane	Chargée de Cours	En poste
28.	DONGMO LEKAGNE Joseph Blaise	Chargé de Cours	En poste
29.	FONKOUA Martin	Chargé de Cours	En poste
30.	KOUOH ELOMBO Ferdinand	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.	OWONA AYISSI Vincent Brice	Chargé de Cours	En poste
35.	Palmer MASUMBE NETONGO	Chargé de Cours	En poste
36.	PECHANGOU NSANGO Sylvain	Chargé de Cours	En poste
37.	WILFRED ANGIE Abia	Chargé de Cours	En poste
38.	MBOUCHE FANMOE Marceline Joëlle	Chargée de Cours	En poste
39.	FOUPOUPOUOGNIGNI Yacouba	Assistant	En poste

2- DÉPARTEMENT DE BIOLOGIE ET PHYSIOLOGIE ANIMALES (BPA) (51)			
1.	AJEAGAH Gideon AGHAINDUM	Professeur	<i>DAARS/FS</i>
2.	BILONG BILONG Charles-Félix	Professeur	Chef de Département
3.	DIMO Théophile	Professeur	En Poste
4.	DJIETO LORDON Champlain	Professeur	En Poste
5.	DZEUFLET DJOMENI Paul Désiré	Professeur	En Poste
6.	ESSOMBA née NTSAMA MBALA	Professeur	<i>Vice Doyen/FMSB/UII</i>
7.	FOMENA Abraham	Professeur	En Poste
8.	KEKEUNOU Sévior	Professeur	En poste
9.	NJAMEN Dieudonné	Professeur	En poste
10.	NJOKOU Flobert	Professeur	En Poste
11.	NOLA Moïse	Professeur	En poste
12.	TAN Paul VERNYUY	Professeur	En poste
13.	TCHUEM TCHUENTE Louis Albert	Professeur	<i>Inspecteur de service Coord Prog/MINSANTE</i>
14.	ZEBAZE TOGOUET Serge Hubert	Professeur	En poste
15.	ALENE Désirée Chantal	Maître de Conférences	<i>Chef Service/MINESUP</i>
16.	BILANDA Danielle Claude	Maître de Conférences	En poste
17.	DJIOGUE Séfirin	Maître de Conférences	En poste
18.	JATSA B. Hermine épouse MEGAPTCHE	Maître de Conférences	En Poste
19.	LEKEUFACK FOLEFACK Guy B.	Maître de Conférences	En poste
20.	MBENOUN MASSE Paul Serge	Maître de Conférences	En poste
21.	MEGNEKOU Rosette	Maître de Conférences	En poste
22.	MONY Ruth épouse NTONE	Maître de Conférences	En Poste
23.	NGUEGUIM TSOFAK Florence	Maître de Conférences	En poste
24.	NGUEMBOCK	Maître de Conférences	En poste
25.	TOMBI Jeannette	Maître de Conférences	En poste
26.	ATSAMO Albert Donatien	Chargé de Cours	En poste
27.	BASSOCK BAYIHA Etienne Didier	Chargé de Cours	En poste
28.	DONFACK Mireille	Chargée de Cours	En poste

29.	ESSAMA MBIDA Désirée Sandrine	Chargée de Cours	En poste
30.	ETEME ENAMA Serge	Chargé de Cours	En poste
31.	FEUGANG YOUMSSI François	Chargé de Cours	En poste
32.	GONWOUO NONO Legrand	Chargé de Cours	En poste
33.	GOUNOUE KAMKUMO Raceline	Chargée de Cours	En poste
34.	KANDEDA KAVAYE Antoine	Chargé de Cours	En poste
35.	KOGA MANG DOBARA	Chargé de Cours	En poste
36.	LEME BANOCK Lucie	Chargé de Cours	En poste
37.	MAHOB Raymond Joseph	Chargé de Cours	En poste
38.	METCHI D. Mireille Flaure épouse Ghoumo	Chargé de Cours	En poste
39.	MOUNGANG Luciane Marlyse	Chargée de Cours	En poste
40.	MVEYO NDANKEU Yves Patrick	Chargé de Cours	En poste
41.	NGOATEU KENFACK Omer Bébé	Chargé de Cours	En poste
42.	NJUA Clarisse Yafi	Chargée de Cours	<i>Chef Div./U. Bamenda</i>
43.	NOAH EWOTI Olive Vivien	Chargée de Cours	En poste
44.	TADU Zephyrin	Chargé de Cours	En poste
45.	TAMSA ARFAO Antoine	Chargé de Cours	En poste
46.	YEDE	Chargé de Cours	En poste
47.	YOUNOUSSA LAME	Chargé de Cours	En poste
48.	AMBADA NDZENGUE GEORGIA E.	Assistante	En poste
49.	FOKAM Alvine Christelle Epse KEGNE	Assistante	En poste
50.	MAPON NSANGO Indou	Assistant	En poste
51.	NWANE Philippe Bienvenu	Assistant	En poste

3- DÉPARTEMENT DE BIOLOGIE ET PHYSIOLOGIE VÉGÉTALES (BPV) (33)			
1.	AMBANG Zachée	Professeur	Chef de Département
2.	BELL Joseph Martin	Professeur	En poste
3.	DJOCGOUE Pierre François	Professeur	En poste
4.	MBOLO Marie	Professeur	En poste
5.	MOSSEBO Dominique Claude	Professeur	En poste
6.	YOUMBI Emmanuel	Professeur	En poste
7.	ZAPFACK Louis	Professeur	En poste
8.	ANGONI Hyacinthe	Maître de Conférences	En poste
9.	BIYE Elvire Hortense	Maître de Conférences	En poste
10.	MAHBOU SOMO TOUKAM. Gabriel	Maître de Conférences	En poste
11.	MALA Armand William	Maître de Conférences	En poste
12.	MBARGA BINDZI Marie Alain	Maître de Conférences	<i>DAAC /UDla</i>
13.	NDONGO BEKOLO	Maître de Conférences	<i>CE / MINRESI</i>
14.	NGALLE Hermine BILLE	Maître de Conférences	En poste
15.	NGODO MELINGUI Jean Baptiste	Maître de Conférences	En poste
16.	NGONKEU MAGAPTCHE Eddy L.	Maître de Conférences	<i>CT / MINRESI</i>
17.	TONFACK Libert Brice	Maître de Conférences	En poste
18.	TSOATA Esaïe	Maître de Conférences	En poste
19.	ONANA Jean Michel	Maître de Conférences	En poste
20.	DJEUANI Astride Carole	Chargé de Cours	En poste
21.	GOMANDJE Christelle	Chargée de Cours	En poste
22.	MAFFO MAFFO Nicole Liliane	Chargée de Cours	En poste
23.	NNANGA MEBENGA Ruth Laure	Chargée de Cours	En poste
24.	NOUKEU KOUAKAM Armelle	Chargée de Cours	En poste
25.	NSOM ZAMBO Annie Claude épouse PIAL	Chargée de Cours	<i>En détachement UNESCO MALI</i>
26.	Godswill NTSOMBOH NTSEFONG	Chargé de Cours	En poste
27.	KABELONG BANAHOU Louis-Paul-R.	Chargé de Cours	En poste
28.	KONO Léon Dieudonné	Chargé de Cours	En poste
29.	LIBALAH Moses BAKONCK	Chargé de Cours	En poste

30.	LIKENG-LI-NGUE Benoit Constant	Chargé de Cours	En poste
31.	TAEDOUNG Evariste Hermann	Chargé de Cours	En poste
32.	TEMEGNE NONO Carine	Chargée de Cours	En poste
33.	MANGA NDJAGA JUDE	Assistant	En poste

4- DÉPARTEMENT DE CHIMIE INORGANIQUE (CI) (31)			
1.	AGWARA ONDOH Moïse	Professeur	Chef de Département
2.	Florence UFI CHINJE épouse MELO	Professeur	<i>Recteur/U. Ngaoundere</i>
3.	GHOGOMU Paul MINGO	Professeur	<i>Ministre Chargé de Miss.PR</i>
4.	NANSEU Njiki Charles Péguy	Professeur	En poste
5.	NDIFON Peter TEKE	Professeur	<i>CT MINRESI</i>
6.	NDIKONTAR Maurice KOR	Professeur	<i>Vice-Doyen Univ. Bamenda</i>
7.	NENWA Justin	Professeur	En poste
8.	NGAMENI Emmanuel	Professeur	<i>DOYEN FS/U. Ngaoundere</i>
9.	NGOMO Horace MANGA	Professeur	<i>Vice Chancellor/UB</i>
10.	ACAYANKA Elie	Maître de Conférences	En poste
11.	EMADACK Alphonse	Maître de Conférences	En poste
12.	KAMGANG YOUBI Georges	Maître de Conférences	En poste
13.	KEMMEGNE MBOUGUEM Jean C.	Maître de Conférences	En poste
14.	KENNE DEDZO GUSTAVE	Maître de Conférences	En poste
15.	KONG SAKEO	Maître de Conférences	En poste
16.	MBEY Jean Aime	Maître de Conférences	En poste
17.	NDI NSAMI Julius	Maître de Conférences	En poste
18.	NEBAH Née NDO SIRI Bridget NDOYE	Maître de Conférences	<i>CT/ MINPROFF</i>
19.	NJIOMOU C. épouse DJANGANG	Maître de Conférences	En poste
20.	NJOYA Dayirou	Maître de Conférences	En poste
21.	NYAMEN Linda Dyorisse	Maître de Conférences	En poste
22.	PABOUDAM GBAMBIE AWAWOU	Maître de Conférences	En poste
23.	TCHAKOUTE KOUAMO Hervé	Maître de Conférences	En poste
24.	BELIBI BELIBI Placide Désiré	Chargé de Cours	<i>Chef Service/ ENS Bertoua</i>
25.	CHEUMANI YONA Arnaud M.	Chargé de Cours	En poste
26.	KOUOTOU DAOUDA	Chargé de Cours	En poste
27.	MAKON Thomas Beauregard	Chargé de Cours	En poste
28.	NCHIMI NONO KATIA	Chargé de Cours	En poste
29.	NJANKWA NJABONG N. Eric	Chargé de Cours	En poste
30.	PATOUOSSA ISSOFA	Chargé de Cours	En poste
31.	SIEWE Jean Mermoz	Chargé de Cours	En Poste

5- DÉPARTEMENT DE CHIMIE ORGANIQUE (CO) (38)			
1.	DONGO Etienne	Professeur	<i>Vice-Doyen / FSE / UYI</i>
2.	NGOUELA Silvère Augustin	Professeur	<i>Chef de Département / UDS</i>
3.	NYASSE Barthélemy	Professeur	En poste
4.	PEGNYEMB Dieudonné Emmanuel	Professeur	Chef de Département <i>Directeur / MINESUP</i>
5.	WANDJI Jean	Professeur	En poste
6.	MBAZOA née DJAMA Céline	Professeur	En poste
7.	Alex de Théodore ATCHADE	Maître de Conférences	<i>Vice-Doyen / DPSAA / UYI</i>
8.	AMBASSA Pantaléon	Maître de Conférences	En poste
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.	KAMTO Eutrophe Le Doux	Maître de Conférences	En poste
13.	KENMOGNE Marguerite	Maître de Conférences	En poste
14.	KEUMEDJIO Félix	Maître de Conférences	En poste
15.	KOUAM Jacques	Maître de Conférences	En poste
16.	MKOUNGA Pierre	Maître de Conférences	En poste
17.	MVOT AKAK CARINE	Maître de Conférences	En poste
18.	NGO MBING Joséphine	Maître de Conférences	<i>Chef de Cellule / MINRESI</i>
19.	NGONO BIKOBO Dominique Serge	Maître de Conférences	<i>C.E.A/ MINESUP</i>
20.	NOTE LOUGBOT Olivier Placide	Maître de Conférences	<i>DAAC/U. Bertoua</i>
21.	NOUNGOU TCHAMO Diderot	Maître de Conférences	En poste
22.	TABOPDA KUATE Turibio	Maître de Conférences	En poste
23.	TAGATSING FOTSING Maurice	Maître de Conférences	En poste
24.	TCHOUANKEU Jean-Claude	Maître de Conférences	<i>Doyen / FS / UYI</i>
25.	YANKEP Emmanuel	Maître de Conférences	En poste
26.	ZONDEGOUNBA Ernestine	Maître de Conférences	En poste
27.	NGNINTEDO Dominique	Chargé de Cours	En poste
28.	NGOMO Orléans	Chargée de Cours	En poste
29.	OUAHOUE WACHE Blandine M.	Chargée de Cours	En poste
30.	SIELINOUE TEDJON Valérie	Chargé de Cours	En poste
31.	MESSI Angélique Nicolas	Chargé de Cours	En poste
32.	TCHAMGOUE Joseph	Chargé de Cours	En poste
33.	TSAMO TONTSA Armelle	Chargé de Cours	En poste
34.	TSEMEUGNE Joseph	Chargé de Cours	En poste
35.	MUNVERA MFIFEN Aristide	Assistant	En poste
36.	NONO NONO Éric Carly	Assistant	En poste
37.	OUETE NANTCHOUANG Judith L.	Assistante	En poste
38.	TSAFFACK Maurice	Assistant	En poste

6- DÉPARTEMENT D'INFORMATIQUE (IN) (22)

1.	ATSA ETOUNDI Roger	Professeur	<i>Chef Div. / MINESUP</i>
2.	FOUDA NDJODO Marcel Laurent	Professeur	<i>Chef Dpt / ENS UYI Chef IGA. / MINESUP</i>
3.	NDOUNDAM René	Maître de Conférences	En poste
4.	TSOPZE Norbert	Maître de Conférences	En poste
5.	ABESSOLO ALO'O Gislain	Chargé de Cours	<i>Sous-Directeur / MINFOPRA</i>
6.	AMINOUE Halidou	Chargé de Cours	Chef de Département
7.	DJAM Xaviera YOUH - KIMBI	Chargé de Cours	En Poste
8.	DOMGA KOMGUEM Rodrigue	Chargé de Cours	En poste
9.	EBELE Serge Alain	Chargé de Cours	En poste
10.	HAMZA Adamou	Chargé de Cours	En poste
11.	JIOMEKONG AZANZI Fidel	Chargé de Cours	En poste
12.	KOUOKAM KOUOKAM E. A.	Chargé de Cours	En poste
13.	MELATAGIA YONTA Paulin	Chargé de Cours	En poste
14.	MONTHE DJIADEU Valery M.	Chargé de Cours	En poste
15.	OLE OLE Daniel Claude Delort	Chargé de Cours	<i>Directeur adjoint ENSET Ebolowa</i>
16.	TAPAMO Hyppolite	Chargé de Cours	En poste
17.	BAYEM Jacques Narcisse	Assistant	En poste
18.	EKODECK Stéphane Gaël Raymond	Assistant	En poste
19.	MAKEMBE S. Oswald	Assistant	En poste
20.	MESSI NGUELE Thomas	Assistant	En poste
21.	NKONDOCK. MI. BAHANACK.N.	Assistant	En poste
22.	NZEKON NZEKO'O Arnel Jacques	Assistant	En poste

7- DÉPARTEMENT DE MATHÉMATIQUES (MA) (31)			
1.	AYISSI Raoult Domingo	Professeur	Chef de Département
2.	EMVUDU WONO Yves S.	Professeur	<i>Inspecteur MINESUP</i>
3.	KIANPI Maurice	Maître de Conférences	En poste
4.	MBANG Joseph	Maître de Conférences	En poste
5.	MBEHOU Mohamed	Maître de Conférences	En poste
6.	MBELE BIDIMA Martin Ledoux	Maître de Conférences	En poste
7.	NOUNDJEU Pierre	Maître de Conférences	<i>Chef Service des Programmes & Diplômes/FS/UYI</i>
8.	TAKAM SOH Patrice	Maître de Conférences	En poste
9.	TCHAPNDA NJABO Sophonie B.	Maître de Conférences	<i>Directeur/AIMS Rwanda</i>
10.	TCHOUNDJA Edgar Landry	Maître de Conférences	En poste
11.	AGHOUKENG JIOFACK Jean Gérard	Chargé de Cours	<i>Chef Cellule MINEPAT</i>
12.	BOGSO ANTOINE MARIE	Chargé de Cours	En poste
13.	CHENDJOU Gilbert	Chargé de Cours	En poste
14.	DJIADU NGAHA Michel	Chargé de Cours	En poste
15.	DOUANLA YONTA Herman	Chargé de Cours	En poste
16.	KIKI Maxime Armand	Chargé de Cours	En poste
17.	MBAKOP Guy Merlin	Chargé de Cours	En poste
18.	MENGUE MENGUE David Joe	Chargé de Cours	<i>Chef Dpt /ENS U. Maroua</i>
19.	NGUEFACK Bernard	Chargé de Cours	En poste
20.	NIMPA PEFOUKEU Romain	Chargée de Cours	En poste
21.	OGADOA AMASSAYOGA	Chargée de Cours	En poste
22.	POLA DOUNDOU Emmanuel	Chargé de Cours	<i>En stage</i>
23.	TCHEUTIA Daniel Duviol	Chargé de Cours	En poste
24.	TETSADJIO TCHILEPECK M. E.	Chargé de Cours	En poste
25.	BITYE MVONDO Esther Claudine	Assistante	En poste
26.	FOKAM Jean Marcel	Assistant	En poste
27.	LOUMNGAM KAMGA Victor	Assistant	En poste
28.	MBATAKOU Salomon Joseph	Assistant	En poste
29.	MBIAKOP Hilaire George	Assistant	En poste
30.	MEFENZA NOUNTU Thiery	Assistant	En poste
31.	TENKEU JEUFACK Yannick Léa	Assistant	En poste

8- DÉPARTEMENT DE MICROBIOLOGIE (MIB) (22)			
1.	ESSIA NGANG Jean Justin	Professeur	Chef de Département
2.	NYEGUE Maximilienne Ascension	Professeur	<i>Vice-Doyen / DSSE/FS/UYI</i>
3.	NWAGA Dieudonné M.	Professeur	En poste
4.	ASSAM ASSAM Jean Paul	Maître de Conférences	En poste
5.	BOUGNOM Blaise Pascal	Maître de Conférences	En poste
6.	BOYOMO ONANA	Maître de Conférences	En poste
7.	KOUITCHEU MABEKE Epse KOUAM Laure Brigitte	Maître de Conférences	En poste
8.	RIWOM Sara Honorine	Maître de Conférences	En poste
9.	SADO KAMDEM Sylvain Leroy	Maître de Conférences	En poste
10.	BODA Maurice	Chargé de Cours	En position d'absence irrégulière
11.	ESSONO OBOUGOU Germain G.	Chargé de Cours	En poste
12.	NJIKI BIKOÏ Jacky	Chargée de Cours	En poste
13.	TCHIKOUA Roger	Chargé de Cours	En poste
14.	ESSONO Damien Marie	Chargé de Cours	En poste

15.	LAMYE Glory MOH	Chargé de Cours	En poste
16.	MEYIN A EBONG Solange	Chargée de Cours	En poste
17.	NKOUDOU ZE Nardis	Chargé de Cours	En poste
18.	TAMATCHO KWEYANG Blandine P.	Chargée de Cours	En poste
19.	TOBOLBAÏ Richard	Chargé de Cours	En poste
20.	MONI NDEDI Esther Del Florence	Assistante	En poste
21.	NKOUÉ TONG ABRAHAM	Assistant	En poste
22.	SAKE NGANE Carole Stéphanie	Assistante	En poste

9. DEPARTEMENT DE PYSIQUE(PHY) (43)			
1.	BEN- BOLIE Germain Hubert	Professeur	En poste
2.	DJUIDJE KENMOE épouse ALOYEM	Professeur	En poste
3.	EKOBENA FOU DA Henri Paul	Professeur	<i>Vice-Recteur / U. Ngaoundéré</i>
4.	ESSIMBI ZOBO Bernard	Professeur	En poste
5.	NANA ENGO Serge Guy	Professeur	En poste
6.	NANA NBENDJO Blaise	Professeur	En poste
7.	NDJAKA Jean Marie Bienvenu	Professeur	Chef de Département
8.	NJANDJOCK NOUCK Philippe	Professeur	En poste
9.	NOUAYOU Robert	Professeur	En poste
10.	PEMHA Elkana	Professeur	En poste
11.	SAIDOU	Professeur	<i>Chef de centre/IRGM/MINRESI</i>
12.	TABOD Charles TABOD	Professeur	<i>Doyen FSUniv/Bda</i>
13.	TCHAWOUA Clément	Professeur	En poste
14.	WOAFO Paul	Professeur	En poste
15.	ZEKENG Serge Sylvain	Professeur	En poste
16.	BIYA MOTTO Frédéric	Maître de Conférences	<i>DG/HYDRO Mekin</i>
17.	BODO Bertrand	Maître de Conférences	En poste
18.	ENYEGUE A NYAM épouse BELINGA	Maître de Conférences	En poste
19.	EYEBE FOU DA Jean sire	Maître de Conférences	En poste
20.	FEWO Serge Ibraïd	Maître de Conférences	En poste
21.	HONA Jacques	Maître de Conférences	En poste
22.	MBINACK Clément	Maître de Conférences	En poste
23.	MBONO SAMBA Yves Christian U.	Maître de Conférences	En poste
24.	NDOP Joseph	Maître de Conférences	En poste
25.	SIEWE SIEWE Martin	Maître de Conférences	En poste
26.	SIMO Elie	Maître de Conférences	En poste
27.	VONDOU DerbetiniAppolinaire	Maître de Conférences	En poste
28.	WAKATA née BEYA Annie	Maître de Conférences	<i>Directeur/ENS/UIYI</i>
29.	ABDOURAHIMI	Chargé de Cours	En poste
30.	CHAMANI Roméo	Chargé de Cours	En poste
31.	EDONGUE HERVAIS	Chargé de Cours	En poste
32.	FOUEDJIO David	Chargé de Cours	<i>Chef Cell. MINADER</i>
33.	MEL'I Joelle Larissa	Chargée de Cours	En poste
34.	MVOGO ALAIN	Chargé de Cours	En poste
35.	WOULACHE Rosalie Laure	Chargée de Cours	<i>Absente depuis Janvier 2022</i>
36.	AYISSI EYEBE Guy François Valérie	Chargé de Cours	En poste
37.	DJIOTANG TCHOTCHOU Lucie A.	Chargée de Cours	En poste
38.	OTTOU ABE Martin Thierry	Chargé de Cours	En poste
39.	TEYOU NGOPOU Ariel	Chargé de Cours	En poste
40.	KAMENI NEMATCHOUA Modeste	Assistant	En poste
41.	LAMARA Maurice	Assistant	En poste
42.	NGA ONGODO Dieudonné	Assistant	En poste
43.	WANDJI NYAMSI William	Assistant	En poste

10- DÉPARTEMENT DE SCIENCES DE LA TERRE (ST) (42)			
1.	BITOM Dieudonné-Lucien	Professeur	<i>Doyen / FASA / UDs</i>
2.	FOUATEU Rose épouse YONGUE	Professeur	En poste
3.	NDAM NGOUPAYOU Jules-Remy	Professeur	En poste
4.	NDJIGUI Paul Désiré	Professeur	Chef de Département
5.	NGOS III Simon	Professeur	En poste
6.	NKOUMBOU Charles	Professeur	En poste
7.	NZENTI Jean-Paul	Professeur	En poste
8.	ABOSSOLO née ANGUE Monique	Maître de Conférences	<i>Vice-Doyen / DRC</i>
9.	BISSO Dieudonné	Maître de Conférences	<i>Directeur/Projet Barrage Memve'ele</i>
10.	EKOMANE Emile	Maître de Conférences	En poste
11.	FUH Calistus Gentry	Maître de Conférences	<i>Sec. D'Etat/MINMIDT</i>
12.	GANNO Sylvestre	Maître de Conférences	En poste
13.	GHOGOMU Richard TANWI	Maître de Conférences	<i>Chef de Département / U. Maroua</i>
14.	MOUNDI Amidou	Maître de Conférences	<i>CT/ MINIMDT</i>
15.	NGO BIDJECK Louise Marie	Maître de Conférences	En poste
16.	NGUEUTCHOUA Gabriel	Maître de Conférences	<i>CEA / MINRESI</i>
17.	NJILAH Isaac KONFOR	Maître de Conférences	En poste
18.	NYECK Bruno	Maître de Conférences	En poste
19.	ONANA Vincent Laurent	Maître de Conférences	<i>Chef service Maintenance & du Matériel / UYII</i>
20.	TCHAKOUNTE J. épouse NUMBEM	Maître de Conférences	<i>Chef.cell / MINRESI</i>
21.	TCHOUANKOUE Jean-Pierre	Maître de Conférences	En poste
22.	TEMGA Jean Pierre	Maître de Conférences	En poste
23.	YENE ATANGANA Joseph Q.	Maître de Conférences	<i>Chef Div. / MINTP</i>
24.	ZO'O ZAME Philémon	Maître de Conférences	<i>DG / ART</i>
25.	ANABA ONANA Achille Basile	Chargé de Cours	En poste
26.	BEKOA Etienne	Chargé de Cours	En poste
27.	ELISE SABABA	Chargé de Cours	En poste
28.	ESSONO Jean	Chargé de Cours	En poste
29.	EYONG JOHN TAKEM	Chargé de Cours	En poste
30.	MAMDEM TAMTO Lionelle Estelle	Chargé de Cours	En poste
31.	MBESSE CECILE Olive	Chargée de Cours	En poste
32.	MBIDA YEM	Chargé de Cours	En poste
33.	METANG Victor	Chargé de Cours	En poste
34.	MINYEM Dieudonné	Chargé de Cours	<i>CD/ U. Maroua</i>
35.	NGO BELNOUN Rose Noël	Chargée de Cours	En poste
36.	NOMO NEGUE Emmanuel	Chargé de Cours	En poste
37.	NTSAMA ATANGANA Jacqueline	Chargé de Cours	En poste
38.	TCHAPTCHET TCHATO De P.	Chargé de Cours	En poste
39.	TEHNA Nathanaël	Chargé de Cours	En poste
40.	FEUMBA Roger	Chargé de Cours	En poste
41.	MBANGA NYOBE Jules	Chargé de Cours	En poste
42.	NGO'O ZE ARNAUD	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	8 (00)	14 (10)	15 (05)	02 (01)	39 (16)
BPA	14 (01)	11 (07)	22 (07)	04 (02)	51 (17)
BPV	06 (01)	10(01)	16 (09)	01 (00)	33 (11)
CI	09(01)	14(04)	08 (01)	00 (00)	31 (06)
CO	06 (01)	20 (04)	08 (03)	04 (01)	38(09)
IN	02 (00)	02 (00)	12 (01)	06 (00)	22 (01)
MAT	02 (00)	08 (00)	14 (01)	07 (01)	31 (02)
MIB	03 (01)	06 (02)	10 (03)	03 (02)	22 (08)
PHY	15 (01)	13 (02)	11 (03)	04 (00)	43 (06)
ST	07 (01)	16 (03)	18 (04)	01 (00)	42(08)
Total	72 (07)	114 (33)	134 (37)	32 (07)	352 (84)

Soit un total de **352 (84)** dont :

- Professeurs **72 (07)**
- Maîtres de Conférences **116 (34)**
- Chargés de Cours **132 (36)**
- Assistants **32 (07)**

() = Nombre de Femmes

DEDICATION

To my wife, Mrs. JIMMAWORK GEBRE GEFERO

ACKNOWLEDGEMENTS

I would like to thank institutions and individuals who helped me in the accomplishment of this thesis work from the beginning to the end. This work would have not been possible without the tremendous support and guidance from all the stakeholders mentioned below.

First and foremost, I would like to thank:

- my Supervisor Professor BELL Joseph Martin for his strong scientific support, useful comments, training me in research methodologies, and guidance from the developmental stages to writing this thesis. Besides, his way of treating foreign students like a father made my life easier; I feel like my homeland is Cameroon;
- my main Advisor and my host institution coordinator for the GENES project, Doctor NGALLE Hermine BILLE, Associate Professor, for guiding and assisting me throughout the whole period of my academic and research study. Without her guidance and constant feedback, my success in my Ph.D. study would not have been achievable;
- my Supervisor and home institution coordinator for the GENES project Doctor WOSENE Gebreselassie Abteu from Jimma University, Ethiopia, for his diligent guidance, supervision, and encouragement from proposal preparation to the final write-up of this thesis. Without his encouragement, insight, and professional expertise, the completion of this work would not have been possible;
- my main Advisor, Dr. CROS David from CIRAD, France for his unreserved and day-to-day monitoring and evaluation from data analysis to the final thesis work. Without his help, this thesis will not stand at this level. Researching under his supervision helped me get much knowledge in the area of plant breeding. His hard-working experience, eagerness, and enthusiasm in his work on genomic selection inspired me to do a lot in the future, generally for Africa and specifically for my country;
- the GENES Project (Mobility for Plant Genomics Scholars to Accelerate Climate-Smart Adaptation Options and Food Security in Africa), an intra-Africa academic mobility project funded by the European Union (EU-GENES) for funding my Ph.D. program;
- the French agricultural research and international cooperation organization (CIRAD) for assigning me a Ph.D. fellowship and giving me all the required materials to undertake my Ph.D. studies;
- PalmElit SAS company for arranging all the required materials to undertake field trials and providing me with phenotypic and molecular data;

- the Faculty of Science of the University of Yaounde I, for giving me this chance to study for my Ph.D. in the Faculty;
- the head of the Department of Plant Biology, Professor AMBANG Zachée, for his intelligent support and guidance right from the day of my arrival in Cameroon by giving a special emphasis on a foreign student like me;
- All the lecturers of the Department of Plant Biology for the lectures, training, and guidance throughout my academic journey in Cameroon;
- the Genetics and Plant Breeding Unit (UGAP) of the Department of Plant Biology of the University of Yaounde I, for making me relax and feeling comfortable during my stay in Cameroon; by following every activity related to my Ph.D. work and by giving their comments and suggestions during my different presentations;
- my academic senior, Doctor NYOUMA Achille and Doctor NTSOMBOH-NTSEFONG Godswill, for thier deep insight of the whole work, correcting and proofreading this thesis work from top to down.

I would also like to thank:

- my wife, Mrs. JIMMAWORK Gebre, who has been extremely supportive of me throughout this entire process of my Ph.D. and has made countless sacrifices to help me get to this point;
- my children, SELHOME Essubalew Getachew, YONAMINE Essubalew Getachew and YOUHAKINE Essubalew Getachew, for continually providing the requisite breaks from philosophy and the motivation to finish my degree with expediency;
- all my parents for their dedication in bringing me up and providing me strong support throughout my life and academic career;
- Repentant father priest Getaye Legesse deserves special thanks for his continued support and encouragement. Without his daily prayers, I doubt that I would be in this position today;
- the Ethiopian community living in both Yaounde and Douala, for their sincere encouragement, support, advice, and exchange of ideas during the entire study period;
- those who helped me directly and indirectly for the betterment of my work but whose names or the institution names are not listed here, may they receive my esteem and deep gratitude;
- all the members of the jury, for their affection towards this research work by agreeing to evaluate and put in their constructive comments and suggestions to improve the quality of this research work.

TABLE OF CONTENTS

DEDICATION	x
ACKNOWLEDGEMENTS	xi
TABLE OF CONTENTS	xiii
LIST OF FIGURES.....	xvi
LIST OF TABLES	xvii
LIST OF ABBREVIATIONS	xviii
LIST OF APPENDICES	xx
ABSTRACT	xxi
RESUMÉ.....	xxii
INTRODUCTION.....	1
CHAPTER I. LITERATURE REVIEW	6
I.1. OIL PALM	6
I.1.1. Taxonomy and morphology of oil palm	6
I.1.2. Types of oil palm in the world.....	11
I.1.2.1. African oil palm, <i>Elaeis guineensis</i> Jacq.....	11
I.1.2.2. American oil palm, <i>Elaeis oleifera</i> HBK Cortes.....	12
I.1.2.3. <i>Elaeis guineensis</i> × <i>Elaeis oleifera</i> Hybrid.....	13
I.1.3. Economic importance and production of oil palm	13
I.2. BREEDING APPROACHES IN OIL PALM.....	15
I.2.1. Mass selection.....	15
I.2.2. Modified reciprocal recurrent selection (MRRS).....	17
I.2.3. Genomic selection	19
I.3. FACTORS AFFECTING THE ACCURACY OF GENOMIC SELECTION	20
I.3.1. Linkage disequilibrium (LD) and effective size (N_e)	21
I.3.2. Marker density and type	22
I.3.3. Traits heritability	23
I.3.4. Statistical models for genomic prediction and trait genetic architecture.....	24
I.3.5. Training and validation population relatedness.....	27
I.3.6. Size and design of the training population	27

I.4. BASIC CONCEPT OF POPULATION GENETICS	28
I.4.1. Linkage disequilibrium.....	29
I.4.2. Effective population size	30
I.4.3. Haplotype sharing.....	31
I.4.4. Fixation index (F_{st})	33
I.5. OIL PALM GENOME MAPPING	33
CHAPTER II. MATERIAL AND METHODS	39
II. 1. MATERIAL.....	39
II.1.1. Basic molecular data	39
II.1.2. Other material.....	42
II.2. METHODS	42
II.2.1. Generation of molecular data	42
II.2.2. Genome mapping	44
II.2.2.1. Construction of the genetic maps.....	44
II.2.2.2. Comparison of genetic and physical maps.....	45
II.2.2.3. Comparative genomics.....	45
II.2.3. Evaluation of genetic diversity of Deli and La Mé populations.....	46
II.2.3.1. Allele and genotype frequencies	46
II.2.3.2. Fixation index (F_{st}).....	46
II.2.4. Estimation of within-population linkage disequilibrium	46
II.2.5. Assessment of haplotype sharing of Deli and La Mé population	47
II.2.6. Determination of the effective population size of Deli and La Mé.....	47
CHAPTER III. RESULTS AND DISCUSSION	49
III.1. RESULTS.....	49
III.1.1. Genetic diversity of Deli and La Mé.....	49
III.1.1.1. Distribution of minor allele and genotype frequencies across the population	49
III.1.1.2. Heterozygosity	50
III.1.1.3. Fixation index (F_{st})	52
III.1.2. Within-population linkage disequilibrium of breeding populations.....	54
III.1.2.1. Persistence phase between Deli and La Mé populations.....	55

III.1.2.2. High-density genetic map	56
III.1.2.3. Comparison of genetic and physical maps	59
III.1.2.4. Comparison between EG5.1 and PMv6 genome sequences.....	62
III.1.3. Haplotype sharing between Deli and La Mé.....	65
III.1.4. Effective size between Deli and La Mé	67
III. 2. DISCUSSION.....	68
III.2.1. Genetic differentiation between Deli and La Mé	68
III.2.1.1. Distribution of minor allele and genotype frequencies across the population	68
III.2.1.2. Heterozygosity	68
III.2.1.3. Fixation index (F_{st})	70
III.2.2. Within-population linkage disequilibrium and persistence phase between Deli and La Mé populations.....	71
III.2.2.1. Within-population linkage disequilibrium.....	71
III.2.2.2. Persistence phase between Deli and La Mé populations	72
III.2.2.3. Comparison of genetic and physical maps	74
III.2.2.4. Comparison between EG5.1 and PMv6 genome sequences.....	76
III.2.3. Haplotype sharing between Deli and La Mé.....	77
III.2.4. Effective size between Deli and La Mé	78
CONCLUSION AND PERSPECTIVES	81
IV.1. CONCLUSION	81
IV.2. RECOMMENDATIONS	82
IV.3. PERSPECTIVES.....	82
REFERENCES.....	84
APPENDICES.....	111
APPENDICES.....	112
PUBLISHED PAPERS	124

LIST OF FIGURES

Fig. 1. Root system of oil palm	7
Fig. 2. Tree morphology of oil palm	8
Fig. 3. Fresh fruits of oil palm-based on exocarp color.	9
Fig. 4. Oil palm fruit types based on endocarp thickness.	10
Fig. 5. Oil palm inflorescences.....	11
Fig. 6. Oil palm tree.	12
Fig. 7. Distinction of different oil palm pollen.	12
Fig. 8. Worldwide distribution of oil palm production.	14
Fig. 9. Scheme of modified reciprocal recurrent selection applied to oil palm.	18
Fig. 10. Diagram of genomic selection (GS) processes.	20
Fig. 11. Formation and development of haplotypes from haploid sequences	32
Fig. 12. Description of the location of plants used.....	40
Fig. 13. Location plan of the 28 trials (GP) of AekLobaTimuer.	41
Fig. 14. Distribution of missing data for oil palm breeding populations per SNP.....	44
Fig. 15. Distribution of minor allele in Deli and La Mé oil palm breeding populations.	49
Fig. 16. Distribution heterozygosity for Deli and La Mé oil palm breeding populations.....	50
Fig. 17. Correlation of heterozygosity per SNPs in oil palm breeding populations	51
Fig. 18. Correlation of alternate allele per SNPs in oil palm breeding populations.....	52
Fig. 19. Fixation index value between oil palm breeding population	53
Fig. 20. Genome-wide pattern of linkage disequilibrium according to cM	54
Fig. 21. Average genome-wide pattern of linkage disequilibrium according to Mbp	54
Fig. 22. Correlation of the r measure of LD between oil palm breeding populations.....	55
Fig. 23. Genetic map of oil palm breeding populations with 4,282 SNP markers.....	57
Fig. 24. Visualization of marker genetic positions versus physical positions.....	60
Fig. 25. Physical map of oil palm breeding populations with 5,598 SNP markers.....	61
Fig. 26 Comparison of the SNP physical positions on the reference genomes.....	64
Fig. 27. Percentage of common haplotypes (in bp) between oil palm breeding populations .	66
Fig. 28. Percentage of common haplotypes (in cM) between oil palm breeding populations	66

LIST OF TABLES

Table I. Oil palm plant material used	41
Table II. Summary of the genetic map	58
Table III. Summary of the physical map	62
Table IV. Percentage of SNPs marker comparison between genome sequences.....	63

LIST OF ABBREVIATIONS

AFLP	:	Amplified Fragment Length Polymorphism
AK	:	Aek Kwasan
ALT	:	Aek Loba Timur
AVROS	:	Algemene Vereniging van Rubber planters ter Oostkust van Sumatra
bp	:	Base pair
BN	:	Bunch Number
BW	:	Bunch Weight
CIRAD	:	Centre de Coopération Internationale en Recherche Agronomique pour le Développement
cM	:	CentiMorgan
CPO	:	Crude Palm Oil
D	:	<i>Dura</i>
DArT	:	Diversity Array Technology
DNA	:	Deoxyribonucleic acid
EG	:	<i>Elaeis guineensis</i>
F _{st}	:	Fixation Index
F/B	:	Fruit per Bunch
GBS	:	Genotyping-by-Sequencing
GBLUP	:	Genomic Best Linear Unbiased Prediction
GEBV	:	Genomic Estimated Breeding Value
GS	:	Genomic Selection
GWAS	:	Genome-Wide Association Study
H ²	:	Broad-sense heritability
h ²	:	Narrow-sense heritability
Kbp	:	Kilobase pair
LD	:	Linkage Disequilibrium
LGs	:	Linkage Groups
LOD	:	Logarithm of the odds
MAF	:	Minor Allele Frequency
MAS	:	Marker-Assisted Selection
Mbp	:	Mega base pair
MRRS	:	Modified Reciprocal Recurrent Selection

N_e	:	Effective size
NIFOR	:	Nigerian Institute for Oil Palm Research
P	:	<i>Pisifera</i>
PCR	:	Polymerase Chain Reaction
PKO	:	Palm Kernel Oil
QTL	:	Quantitative Trait Locus
RAD	:	Restriction Associated DNA tagging
RAPD	:	Random Amplified Polymorphic DNA
RFLP	:	Restriction Fragment Length Polymorphism
RKHS	:	Reproducing Kernel Hilbert Spaces
RRBLUP	:	Random Regression Genomic Best Linear Unbiased Prediction
RRS	:	Reciprocal Recurrent Selection
SEA	:	South-East Asia
SNP	:	Single Nucleotide Polymorphism
SOCFINDO	:	Société Financière des Caoutchoucs d'Indonésie
SPET	:	Single Primer Enrichment Technology
SSR	:	Simple Sequence Repeat
T	:	<i>Tenera</i>
WA	:	West Africa

LIST OF APPENDICES

Appendix 1. Principal fatty acid compositions of the nine major globally traded vegetable oils.	112
Appendix 2. Food and industrial importance of oil palm.	113
Appendix 3. Summary of linkage map constructed in oil palm.	114
Appendix 4. Major steps of genotyping-by-sequencing (GBS) protocol used in plant breeding.	115
Appendix 5. Generation of SNPs working.	116
Appendix 6. Objectives and corresponding published papers.	117
Appendix 7. The logical framework of the specific objective I: evaluation of the genetic diversity between Deli and La Mé oil palm breeding populations.	118
Appendix 8. The logical framework of the specific objective II: estimation within-population linkage disequilibrium for Deli and La Mé oil palm breeding populations.	119
Appendix 9. The logical framework of the specific objective III: assessment of the haplotype sharing between Deli and La Mé oil palm breeding populations.	120
Appendix 10. The logical framework of the specific objective IV: evaluation of effective population size between Deli and La Mé.	121
Appendix 11. Genetic map with 4,759 SNP markers on 16 linkage groups (LG). The y axis indicates the distances in centiMorgan (cM).	122

ABSTRACT

Oil palm (*Elaeis guineensis* Jacq.) is the most efficient oil-bearing crop grown in humid tropical parts of the world. A better understanding of the Genomic Selection (GS) results in the populations involved needs a detailed study of their genome properties. This study aimed to characterize the genome properties of two complex oil palm breeding populations, i.e, Deli and La Mé. The present study considered 423 Deli, 140 La Mé, and 380 Deli × La Mé hybrid crosses with a total of 943 genotyped individuals. A total of 7,324 SNPs, including 5, 598 SNPs located on the anchored sequences of the genome, were involved. The LepMAP3 software was used to construct the genetic linkage map. Analyses of linkage disequilibrium (LD) in cM and in Mbp were performed using the PLINK software. Haplotypes sharing and minor allele frequency (MAF) were analyzed. The effective size (N_e) was estimated using the NeEstimator 2.1 software and pairwise of fixation index (F_{st}) was calculated using the SNPRelate R package. The genetic linkage map constructed included 4, 252 SNPs and spanned 1,778.52 cM, with an average recombination rate of 2.85 cM/Mbp. The LD at $r^2 = 0.3$, considered the minimum to get reliable genomic selection (GS) results, spanned over 1.05 cM/0.22 Mbp in Deli and 0.9 cM/0.21 Mbp in La Mé. The LD decay was faster for Deli than for La Mé. The significant degree of differentiation existing between Deli and La Mé was confirmed by the high F_{st} value (0.53), the pattern of correlation of SNP heterozygosity and allele frequency among populations, as well as the decrease of persistence of LD and of haplotype sharing among populations with increasing SNP distance. The two populations had low N_e (< 5) although a lower N_e was observed in Deli than in La Mé. In conclusion, the study helped to determine the number of markers to be used for future GS studies in oil palm, showing that 10,000 SNPs would be enough to reach the r^2 value of 0.3 in Deli and La Mé. Overall, the results showed strong genetic differentiation between Deli and La Mé, but the level of resemblance between them over short genomic distances likely explained the superiority of GS models ignoring the parental origin of marker alleles over models taking this information into account. Future studies in oil palm should consider population-specific genetic maps, new reference genomes and other breeding populations.

Keywords: Genome properties, genomic selection, hybrid performance, *Elaeis guineensis*, single nucleotide polymorphisms.

RESUMÉ

Le palmier à huile (*Elaeis guineensis* Jacq.) est la culture oléagineuse la plus efficace dans les régions tropicales humides du monde. Une meilleure compréhension des résultats de la sélection génomique (GS) dans les populations concernées nécessite une étude détaillée des propriétés de leur génome. Cette étude visait à caractériser les propriétés génomiques de deux populations complexes de palmiers à huile, à savoir Deli et La Mé. La présente étude a pris en compte 423 Deli, 140 La Mé et 380 croisements hybrides Deli × La Mé avec un total de 943 individus génotypés. Un total de 7324 SNP, dont 5598 SNP situés sur les séquences ancêtres du génome, ont été impliqués. Le logiciel LepMAP3 a été utilisé pour construire la carte de liaison génétique. Des analyses du déséquilibre de liaison en cM et en Mbp ont été réalisées à l'aide du logiciel PLINK. Le partage des haplotypes et la fréquence des allèles mineurs ont été analysés. La taille effective a été estimée à l'aide du logiciel NeEstimator 2.1 et l'indice de fixation par paire a été calculé à l'aide du progiciel SNPRelate R. La carte de liaison génétique construite comprenait 4252 SNP et s'étendait sur 1 778.52 cM, avec un taux de recombinaison moyen de 2.85 cM/Mbp. Le DL à $r^2 = 0.3$, considéré comme le minimum pour obtenir des résultats de sélection génomique fiables, s'étendait sur 1.05 cM/0.22 Mbp à Deli et 0.9 cM/0.21 Mbp à La Mé. La décroissance du DL était plus rapide pour Deli que pour La Mé. Le degré important de différenciation existant entre Deli et La Mé a été confirmé par la valeur élevée de F_{st} (0.53), le modèle de corrélation de l'hétérozygotie SNP et de la fréquence des allèles entre les populations, ainsi que la diminution de la persistance du DL et du partage des haplotypes entre les populations avec l'augmentation de la distance SNP. Les deux populations avaient un N_e faible (< 5), bien qu'un N_e plus faible ait été observé à Deli qu'à La Mé. En conclusion, l'étude a permis de déterminer le nombre de marqueurs à utiliser pour les futures études de GS chez le palmier à huile, en montrant que 10, 000 SNP seraient suffisants pour atteindre la valeur r^2 de 0.3 à Deli et à La Mé. Dans l'ensemble, les résultats ont montré une forte différenciation génétique entre Deli et La Mé, mais le niveau de ressemblance entre eux sur de courtes distances génomiques a probablement expliqué la supériorité des modèles GS ignorant l'origine parentale des allèles des marqueurs par rapport aux modèles prenant en compte cette information. Les études futures sur le palmier à huile devraient prendre en compte les cartes génétiques spécifiques aux populations, les nouveaux génomes de référence et d'autres populations de sélection.

Mots clés : Propriétés du génome, sélection génomique, performances hybrides, *Elaeis guineensis*, polymorphismes mononucléotidiques.

INTRODUCTION

INTRODUCTION

Oil palm (*Elaeis guineensis* Jacq.) is a perennial tropical oil-producing crop that belongs to the family of *Arecaceae* and the genus *Elaeis* (Hartley, 1988; Ithnin & Din, 2020). It originated from the tropical West/Central African coastal belt between Guinea/Liberia and northern Angola (Hartley, 1977, 1988; Corley & Tinker, 2015). It is naturally cross-pollinated, monoecious, allogamous, and diploid with a chromosome number of $2n = 2x = 32$ and a genome sequence of 1.8 gigabases (Singh *et al.*, 2013; Corley & Tinker, 2015; Ithnin & Din, 2020). The economic life span of oil palm ranges from 25 to 30 years and it is mainly cultivated in humid tropical zones of Asia, Africa, and the Americas, from where its products are exported to global markets (Barcelos *et al.*, 2015; Corley & Tinker, 2015; Murphy *et al.*, 2020).

Oil extracted from oil palm is classified into Crude Palm Oil (CPO) which is produced from the fibrous mesocarp, and Palm Kernel Oil (PKO) produced by the kernel (Mba *et al.*, 2015). The former generated oil with a dark orange-red, semi-solid fluid, whilst the latter produced oil with a white-yellow oil that is primarily derived from the endosperm tissue of the kernel (seed). Generally, about 89% of the total fruit oil in palm trees is obtained from the mesocarp, and the remaining 11% comes from the seed (Murphy *et al.*, 2020). CPO contains both healthy and beneficial substances including triacylglycerols (TAGs), vitamin E, carotenoids, and phytosterols as well as impurities such as phospholipids, free fatty acids (FFAs), gums, and lipid oxidation products (Mancini *et al.*, 2015). PKO contains a high composition of unsaturated fatty acids such as oleic and linoleic, whereas CPO contains more saturated fatty acids and lauric acid (Nainggolan & Sinaga, 2021). The oil obtained from oil palm has better balanced fatty acid compositions than the major globally traded vegetable oils (Appendix 1) i.e., Soybean, Rapeseed, Sunflower, Peanut, Cottonseed, Coconut, and Olive (Murphy *et al.*, 2020).

The total world vegetable oil production is currently around 200 million metric tons (MT), led by palm oil (75 MT), followed by soybean oil (60 MT), rapeseed oil (28 MT), and sunflower oil (19 MT) (Statista, 2021). It supplies 40% of the total traded vegetable oil and globally produces an annual 81 MT of oil on about 23 million hectares (Murphy *et al.*, 2020; Yue *et al.*, 2021). Oil palm produces an average oil yield of 4 metric tons/ha/yr, which is approximately 10 times higher than soybean (Babu & Mathur, 2016; Corley & Tinker 2016).

Despite its leading position in the vegetable oil market, oil palm production is still far from its potential due to several biotic and abiotic constraints. Climate change, land, labor shortage and diseases (in particular vascular wilt, ganoderma and bud rot) are the major factors that hinder the yield and quality of palm oil across the world (Corley, 2009; Barcelos *et al.*, 2015; Kwong *et al.*, 2016; Pirker *et al.*, 2016; Murphy *et al.*, 2020).

As oil palm is a multipurpose crop (Appendix 2) and the cheapest source of vegetable oil and fat available in the world (Murphy, 2014), its global cultivation has some challenges and controversy despite its huge economical importance. Due to the large area requirements, the cultivation brings environmental and ecological impacts which resulted in significant habitat loss, and reductions in biodiversity in complex ecosystems which increase greenhouse gas (GHGs) emissions (Barcelos *et al.*, 2015; Meijaard *et al.*, 2020). Conversely, habitat fragmentation and increased pollution brought on by peat soil burning to make way for new plantations and promote deforestation might further enhance the release of GHGs that contribute to climate change (Cook *et al.*, 2018; Dislich *et al.*, 2017; Tonks *et al.*, 2017). Further, the expansion increases CO₂ emissions by 6-17% due to forest loss (Wich *et al.*, 2012). Equally important, health and related issues due to the high consumption of oil palm are also other controversial issues for oil palm diversification (Qaim *et al.*, 2020). Because palm oil is becoming more widely available and consumed, together with the fact that it contains a lot of saturated fatty acids, it is believed to be a factor in the rise in cancer and cardiovascular disease rates which resulted in high mortality rates increased (Chen *et al.*, 2011; Mancini *et al.*, 2015; Ismail *et al.*, 2018; Kadandale *et al.*, 2019). Moreover, the steady growth of the world population is expected to reach 9-11 billion by 2050 (Röös *et al.*, 2017), and the world demand for vegetable oils, by the same year, is estimated to reach 240–250 Mt (Babu *et al.*, 2021). Just over 300% of this demand will have to be met by palm oil. Therefore, before reaching the aforementioned year, it is crucial to apply new oil palm breeding strategies that aim for considerably better yielding varieties, improved oil profiles enhanced disease resistance, and environmentally friendly (Murphy *et al.*, 2020; Rival, 2017).

Genetic improvement in oil palm has been done through both conventional breeding methods (i.e, mass selection and modified reciprocal recurrent selection (MRRS)) and modern biotechnological approaches (tissue culture, genetic modification, and marker-assisted selection (MAS)) (Murphy *et al.*, 2020; John Martin *et al.*, 2022). Breeding through mass selection helped to get the current breeding populations grouped into two complementary

groups (A and B) based on the characteristics of their bunch production (Nyouma *et al.*, 2019). MRRS was also used to help exploit the hybrid vigor for bunch production that appeared in the crosses (A × B), and they enabled better estimates of genetic values than mass selection (Nyouma *et al.*, 2019; Soh *et al.*, 2017). However, oil palm breeding through the conventional breeding methods showed many constraints due to its costly, time-consuming, large size area needed, long breeding cycle, and limited number of tested individuals (Wong & Bernardo, 2008; Cros *et al.*, 2014; Jin *et al.*, 2016; Seng *et al.*, 2016). Fortunately, the introduction of novel genotypes for breeding through biotechnological technology has revolutionized traditional plant breeding methods (John Martin *et al.*, 2022). Therefore, to provide a solution while ensuring a sustainable future, marker-assisted breeding has been introduced into oil palm breeding programs (Soh *et al.*, 2017).

Genomic selection (GS) is a highly effective MAS method that helps to improve quantitative traits, especially yield (Meuwissen *et al.*, 2001). It is a MAS method with a high density of markers on the entire genome so that at least one marker can be in LD with each quantitative trait locus (QTL) (Goddard & Hayes, 2007). Unlike QTL-based MAS, GS utilizes dense genome-wide markers simultaneously, to predict the genetic values of individuals in the selection population (Grattapaglia *et al.*, 2018; Wang *et al.*, 2018). It is one of the most effective MAS methods used to improve quantitative traits (Heffner *et al.*, 2009). Studies on the application of GS in oil palm brought positive results. Thus, GS could improve oil palm clonal selection (Nyouma *et al.*, 2020) and the selection of parents to use for hybrid crossings (Cros *et al.*, 2017, Nyouma *et al.*, 2022). The benefit of GS in oil palm relies on its ability to enhance selection intensity and/or shorten the generation interval, thus increasing the annual genetic gain (Nyouma *et al.*, 2019). Different studies carried out provided a significant amount of information concerning the conditions of implementation of GS in this species. For example, in Deli and La Mé, GS has been implemented with relatively small training populations (< 150) and low marker density (< 2,000) Cros *et al.* (2017); Nyouma *et al.* (2020) and models ignoring the parental origin of marker alleles were found to be more accurate than models accounting for this information (Nyouma *et al.*, 2020, 2022). To better understand GS results in populations involved, an in-depth study of their detailed genome properties should be conducted. However, to the knowledge, such research has not yet been done in oil palm, particularly regarding LD, N_e , haplotype sharing, and fixation index (F_{st}), factors known to affect GS accuracy.

Indeed, these parameters affect GS accuracy, and understanding their effect when predicting the performances of oil palm hybrids would help to define a more efficient and robust GS scheme for this species. In particular, this could help to understand the variations in GS accuracy observed among families of hybrid parents, thus helping in choosing application families given the available training population. However, the knowledge of the impact of genome properties in oil palm hybrid variety development is less compared to the case of other plants and animals, for instance, hybrid cultivars development in Maize, Cattle, and Pig Liu *et al.* (2018); Zhang *et al.* (2019), Black spruce Lenz *et al.* (2017), Barley (Zhong *et al.*, 2009). Likewise, there is also no study on the genome properties of Deli and La Mé using genome-wide markers.

The **hypothesis of the research** adopted here assumes that the use of genome-wide markers among families of hybrid parents **may not** significantly increase the knowledge of the genome properties of Deli and La Mé oil palm parental populations used for hybrid breeding. A few **research questions** emerge from this hypothesis.

- What is the level of genetic diversity between Deli and La Mé oil palm breeding populations using genome-wide markers?
- What is the impact of genome properties on Deli and La Mé oil palm breeding populations for future GS hybrid variety development studies?
- How many SNPs are required for MAS and are enough for genomic predictions in oil palm hybrid variety development?
- Why across-population SNPs GS model are better than the population-specific effects SNPs alleles GS model?

The general objective of this study is to characterize the genome properties of Deli and La Mé oil palm breeding populations through genome-wide markers for better palm oil yield. The specific objectives are:

- to evaluate the genetic diversity between Deli and La Mé oil palm breeding populations;
- to estimate within-population linkage disequilibrium between Deli and La Mé oil palm breeding populations;
- to assess their haplotype sharing between Deli and La Mé oil palm breeding populations;

- to determine their effective population size for Deli and La Mé oil palm breeding populations.

This introductory section is immediately followed by chapter I, a review of the literature which will present, not only the target plant (oil palm) but also the fundamental concepts inherent to the problematic of the research. After this first chapter, the document will go along with chapter II (material and methods) and chapter III (results and the discussion). A conclusion, recommendations, and perspectives will close the presentation of the study.

CHAPTER I. LITERATURE REVIEW

CHAPTER I. LITERATURE REVIEW

I.1. OIL PALM

I.1.1. Taxonomy and morphology of oil palm

The word *Elaeis* has its origin in the Greek word *Elaion*, meaning oil (Nair, 2021). It belongs to the family of *Arecaceae* and the genus *Elaeis*. The species name of the oil palm i.e., *guineensis*, implies the plant originated from the Gulf of Guinea. Jacquin in 1763 gave the scientific name for oil palm as *Elaeis guineensis* Jacq. (Purseglove, 1986; Hartley, 1988). The *Arecaceae* are placed in the order *Arecales* and grouped with *Cocos* and other genera in the subfamily *Cocosoideae* and tribe *Cocoeae* (Corley & Tinker, 2003). It consists of two species, the *E. guineensis* native to Africa, and *E. oleifera* (Kunth) Cortes indigenous to South and Central America and the third one is *E. odora* (or *Barcella odora*), although its taxonomy has not been confirmed (Corley & Tinker, 2003; Soh *et al.*, 2003). Currently, *E. guineensis* is the major economic oil-producing species, and *E. oleifera* has a much lower oil content and is used only locally in their natural area of distribution (Corley & Tinker, 2015). The *Arecoideae* subfamily of *E. guineensis* is the most diverse and largest of the five subfamilies in the *Arecaceae* family, which contains around 60% of palm genera that is almost equal to 107 out of 183 and greater than 50% of species, approximately 1300 out of 2400 (Baker *et al.*, 2011).

Healthy trees of oil palm roots produced primary, secondary, tertiary, and quaternary roots (Hartley, 1988). The primary roots grow up to 5–10 mm in diameter and extend downwards to the base of the palm or more or less in an outwards horizontal direction. From them, the secondary roots develop growing up to 1–4 mm in diameter in both downward and upwards directions. From the secondary roots, the tertiary roots originated and grow up to 0.5-1.5 mm in diameter and 20 cm in length. The tertiary roots give rise to quaternary roots, growing up to 3 cm in length and only 0.2-0.5 mm in diameter (Corley & Tinker, 2003). Generally, oil palm produced an adventitious root system that arises from the root plate, and after turning from the juvenile phase, it produced eight different morphological types of roots based on their development pattern and state of differentiation namely, primary vertical and horizontal roots, secondary horizontal roots, upward growing secondary vertical roots and downward growing secondary vertical roots, superficial and deep tertiary roots and quaternary roots (Jourdan & Rey, 1997; Intara *et al.*, 2018). The root system of the oil palm is depicted in Fig. 1 and outlines that there are no root hairs (Jourdan & Rey, 1997).



Fig. 1. Root system of oil palm (Intara *et al.*, 2018).

Oil palm produced a wide stem base, after the seedling stage without the internodal elongation with very little increment in the first three years (Fig. 2a) (Corley & Tinker, 2003). The stem or stipe of the oil palm is erect, cylindrical, solitary, scarred, unbranched, and heavily ringed with short internodes. It is covered with petiole bases in young palms, and smooth in older trees (>10–12 years old). It reaches a height between 15 and 30 m with varying diameters irrespective of the genetic origin and growing climatic conditions and the stem can last up to 300 years (Nair, 2010). It facilitates the transportation and storage of nutrients. In commercial oil palm plantation farms, the growth of the stem is restricted when it exceeds 15 m in height, 80 to 110 cm in diameter at the base, and 40-45 cm on the cylindrical area (Fig. 2b). If the palm, grows beyond this height it will create a problem for harvesting and maintenance. So, oil palm should be removed 25-30 years after planting (Hartley, 1977).

Oil palm leaf development starts in an adult stage at the crown with the palm prolonged up to leaf buds or primordia separating laterally from the apical meristem. The adult palm produced 40-60 leaves within the apical bud, each 5-9 m long, weighing 5-8 kg and this leaf persists as a strong fibrous sheet. At the mature stage, palm leaf is pinnate, bearing linear leaflets or pinnae on each side of the leaf stalk (Fig. 2c). The oil palm-produced pinnate leaf blade has a pair of leaflets with strong central nerves, specifically at the base, and is green on both surfaces. Healthy palm trees at the age of 20 to 40 years, carry 190 to 200 leaflets with a length of 70 to 90cm. It produced wide saw-toothed petioles, with an average length of 70 cm to 1.10 m and 25 cm wide with varying color ranges from green, yellowish-green, or yellow-ochre, and darker central stripe color. The leaflets are arranged on two lateral planes. It produced a hard and fibrous leaf stalk and it grows up to 8 m. Annually the palm produced 30 and 40 leaves at the age of 2-4 years of age, then after the production gradually declines and reaches 20-25 per annum from about 8 years onwards (Corley & Tinker, 2003; Nair, 2010).



Fig. 2. Tree morphology of oil palm. (a) Oil palm seedling; (b) Palm trees on a plantation; (c) Crown with bunches (Godswill *et al.*, 2016).

Oil palm produced a sessile fruit drupe with different shapes ranging from nearly spherical to ovoid or elongated and bulging somewhat at the top. The length of fruits varies from 2 cm up to more than 5 cm, from 3 g to over 30 g of weight (Corley & Tinker, 2003). Oil palm produced either *nigrescens* or *virescens* fruit type (Fig. 3c). The former is characterized by the production of unripe fruits with deep-violet to black at the fruit apex and underwent minimal color change upon ripening, becoming red at the base when ripe (Fig. 3a). While the latter is produced in unripe fruits with green color at the apex of the fruit and changed to reddish-orange upon ripening (Fig. 3b) (Singh *et al.*, 2014).

The fruit pericarp of the palm has three major layers, i.e, the outer exocarp or skin, the mesocarp or pulp, and the endocarp or shell. The mesocarp of the fruit gives the palm oil surrounding a nut with a hard shell. This shell covers the palm kernel (seed) which gives PKO and residual palm kernel cake used as food for livestock. Irrespective of the various factor the oil content of the mesocarp of ripe fruit varies from under 40% to over 60% (Corley & Tinker, 2003). Oil palm produces a fresh fruit bunch varying in size and weight from the development of anthesis until 100 days or more after anthesis. In most cases, oil formation in the kernel begins around 70 days after anthesis development and is probably complete by about 120 days (Corley & Tinker, 2003). Then, the augmentation of oil in the mesocarp starts 18 to 22 weeks after pollination, and this remains under production until the fruit is overripe (Teh *et al.*, 2014).



Fig. 3. Fresh fruits of oil palm-based on exocarp color. (a) Nigrescens (Nig) fruits; (b) Virescens (Vir) fruits; (c) ripe nigrescens and virescens fruit bunches (Singh *et al.*, 2014).

E. guineensis is also classified based on the shell thickness which is a key factor for the genetic improvement of oil palm. Depending on this trait, there are three types of oil palm (Fig. 4), the wild type, *dura* (Sh^+Sh^+) whose fruit has a thick shell, the mutant type *pisifera* (Sh^-Sh^-) with shelled fruit and the hybrid type *tenera* (Sh^+Sh^-) issued from the cross between *dura* and *pisifera* which shell fruit is thin (Montoya *et al.*, 2013). The thickness of the endocarp also influences the oil content of the fruits of the three palm types. Based on this, *dura* consists of 15%, *pisifera* contains 25% and *tenera* consists of 36% of the oil in the fruit (Babu & Mathur, 2016).

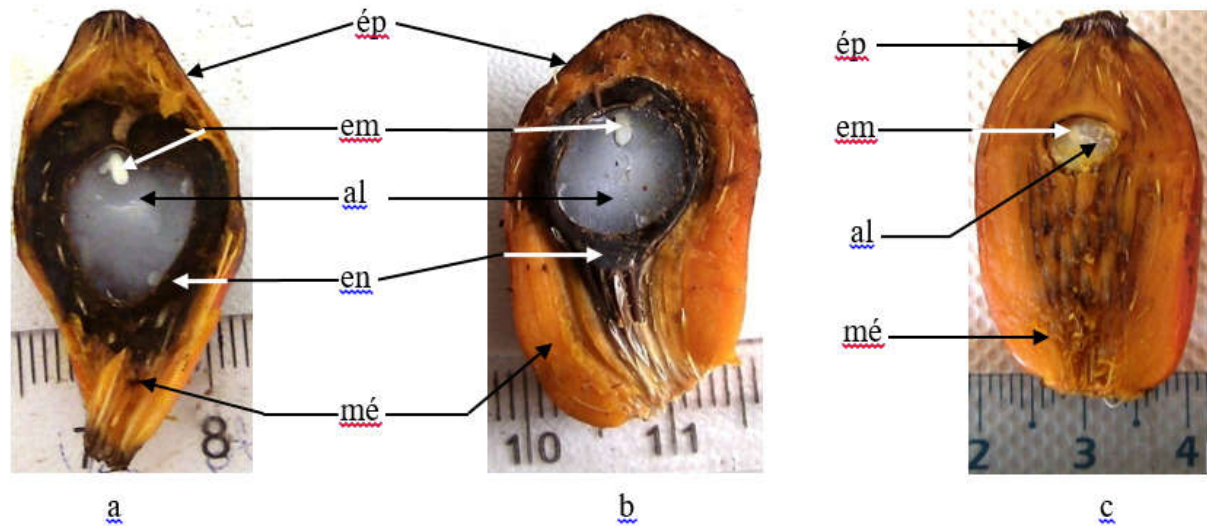
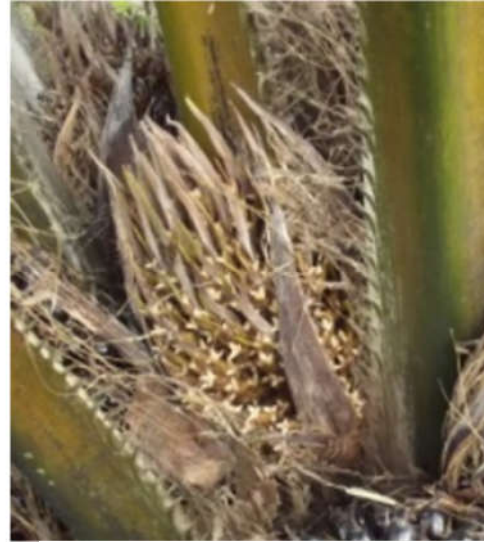


Fig. 4. Oil palm fruit types based on endocarp thickness.
 (a) *dura*, (b) *tenera* and (c) *pisifera* (Ngalle, 2016)

The oil palm usually produces a monoecious flower with male and female flowers existing separately on the same plant, which results in an allogamous mode of reproduction. However, very often it produces also mixed inflorescences in the axils of the leaves. An inflorescence originated from the leaf axils, with male and female but, not both inflorescences per leaf axil and some of them abort before the emergence (Corley & Tinker, 2003; Godswill *et al.*, 2016). Oil palm potentially produced both male and female reproductive organs with a very rare case producing androecium and gynoecium flowers that give rise to a hermaphrodite flower (Hartley, 1988). In most cases, young oil palm trees produced mixed inflorescences, with both male and female spikelets (Corley & Tinker, 2003). In the palm tree, the male flower exists close to the trunk on short pedicels. Whereas, the female flowers are set in clusters close to the trunk on short heavy pedicels. Unlike male flowers, female flowers are produced with large clusters of stalks on short heavy pedicels. Before the development of anthesis, the female inflorescence reaches a length of 30 cm or more. Both flowers comprise a central stem holding about 200 flower-bearing spikelets (Fig. 5). The spikelet of the male inflorescence contains about 1000 flowers while, the female inflorescence contains 15-30 flowers (Godswill *et al.*, 2016). In oil palm, at the time of the rainy season, all the oil palm flowers found in the spikelet are open within two days and it lasting up to 4 days. Before the opening of each flower, the sessile flower is completely enclosed by a triangular bract. Male flowers anthers dehiscence by vertical slits (Corley & Tinker, 2003). Irrespective of various factors individual palm trees produced an average of 10 ± 2.5 male inflorescences and 7 ± 2 female inflorescences each year. On average the female inflorescence weighs about 8 kg (Tandon, 2001).



(a) Male inflorescence



(b) Female inflorescence

Fig. 5. Oil palm inflorescences (Godswill *et al.*, 2016).

I.1.2. Types of oil palm in the world

The African oil palm, *Elaeis guineensis* Jacq. and the American oil palm *Elaeis oleifera* HBK Cortes, are the two most economically and genetically important oil palm species under the genus *Elaeis*. This genus also has two less economically important species namely *E. melanococca* and *E. madagascariensis*, which are used for the genetic improvement of *E. guineensis* Jacq. *E. odora* syn. *Barcella odora* is often cited as belonging to the genus *Elaeis* (Jacquemard *et al.*, 2001).

I.1.2.1. African oil palm, *Elaeis guineensis* Jacq.

Elaeis guineensis Jacq. and/or the African oil palm (Fig. 6) is also known as the olive tree of Guinea, due to its origin in the highlands of the Fouta Djallon district of the Gulf of Guinea, from which the name *guineensis* was given. It is a diploid plant ($2n = 2x = 32$) and monocotyledon, which belongs to the *Arecaceae* family (Corley & Tinker, 2016; Ithnin & Din, 2020).



Fig. 6. Oil palm tree (Jalani *et al.*, 1997).

Globally the African origin of *E. guineensis* was controversial by the international oil palm researcher until Zeven (1964) confirmed that it originated from Africa. His finding indicated that the pollen of *E. guineensis* was found in the young tertiary sediments of the Miocene in the Niger Delta, which typically resembles the present-day oil palm (Fig. 7).

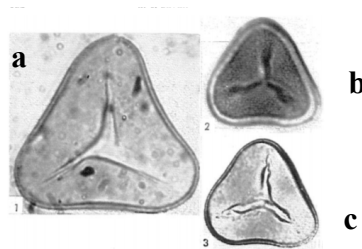


Fig. 7. Distinction of different oil palm pollen. (a) Fossil pollen seemingly related to a fern spore (Nigeria) $\times 1750$; (b), (c) Fresh pollen of oil palm *pisifera* (Nigeria) $\times 1300$ (Zeven, 1964).

I.1.2.2. American oil palm, *Elaeis oleifera* HBK Cortes

Elaeis oleifera HBK Cortes also known as the American oil palm which resembles *E. madagascariensis* Hartley (1988) is native to tropical Latin America, extending from Mexico in the north to the Amazonas to Bolivia, Brazil, Colombia, and Peru in the south and along the Pacific and Atlantic coasts (Corley & Tinker, 2016; Ithnin *et al.*, 2017). It is also found in the coppice in open grasslands and wetland areas and spreading exits in association with the indigenous Indian migratory movements (Soh *et al.*, 2003). The oil obtained from *Elaeis oleifera* is small in quantity as compared to *Elaeis guineensis* Jacq. However, *E oleifera* has numerous breeding advantages due to its unique characteristics, for instance, good oil quality, slow growth rate, and short-spined bunch's resistance to fatal yellowing; though with low fruit-

to-bunch ratio (Corley & Tinker, 2003; Ithnin *et al.*, 2017). It produced numerous fruits in a parthenocarpic method and the oil obtained from its pulp is highly characterized by a higher content of unsaturated fatty acids, which gives rise to fluidity similar to that of olive oil (Meunier, 1969; Vossen, 1974; Ithnin *et al.*, 2017).

I.1.2.3. *Elaeis guineensis* × *Elaeis oleifera* Hybrid

The interspecific hybridization between *E. guineensis* (G) and *E. oleifera* (O), i.e, G×O hybrids, has frequently been used at the experimental level to obtain a hybrid having more important traits than the parents (Meunier & Hardon, 1976). In this regard, several scientific research findings indicate that the hybrid has pros and cons over their parents, for instance, screening partial resistance to the bud rot disease from the O×G hybrid (Hormaza *et al.*, 2012). Regarding the improvement of oil quality, Cadena *et al.* (2013) reported the G×O hybrids to have less lipase activity and higher iodine value than *E. guineensis*. Hybridization between the two oil palm species substantially modifies the biosynthesis of fatty acids (Mozzon *et al.*, 2013). An improvement in antioxidant capacity from the hybrid compared to their parents was shown (Rodríguez *et al.*, 2016; Ojeda *et al.*, 2017). In most cases for vegetative and yield traits, the hybrid has intermediate characteristics to their parents (Hardon, 1969). The hybrid also has a low percentage of pollen viability and germination due to the inflorescences being less attractive for pollinator insects which resulted in a poor fruit set and yield than the parents (Hardon & Tan, 1969).

I.1.3. Economic importance and production of oil palm

The oil palm is one of the most important oil-producing vegetable crops which is demanded and grown around the globe (Murphy *et al.*, 2020). It is one of the major food security crops in countries suffering from food insecurity by providing rural income and food (Rosas Urióstegui *et al.*, 2018). Nowadays, palm oil is a staple cooking oil that is frequently used in the preparation of food in Africa and Asia, where it is consumed by at least three billion people worldwide (Murphy *et al.*, 2020). It has several economic importance and is much required by various industries; particularly in the food industry. The oil obtained from oil palm cascades into two major groups food industry (with over 80% of the market) and the rest of the chemical industry for the formulation of paints, lipstick, inks, shampoo, chocolate resins, varnishes, plasticizers, biodiesel production (Appendix 2), etc. (Corley, 2009; Montoya *et al.*, 2013; Soh *et al.*, 2017). Generally, it is difficult to find substitutes for palm oil. Despite the current climate change constraints (Barcelos *et al.*, 2015; Pirker *et al.*, 2016; Woittiez *et al.*, 2017), outlined

there is an expansion of the oil palm-based industry in the tropical areas of Africa, Asia, and America due to its high oil production (Corley, 2009; Murphy *et al.*, 2020). The oil harvested from oil palm trees produced a potential oil yield capacity of 18.2 tons/ha/year, which varied between 2-6 t/ha of oil production among different oil-producing crops and which is equivalent to ten times the average potential of oil yield hectare⁻¹ year⁻¹ as compared to other oil-producing crops (like, soybean), this means oil palm requires ten times less land than the other three major oil-producing crops (Babu & Mathur, 2016; Tapia *et al.*, 2021; Yue *et al.*, 2021).

Currently, oil palm is grown in at least 30 countries of the world. It supplies about 40% of all traded vegetable oil (Murphy *et al.*, 2020). World oil palm production is distributed in many countries on different continents (Fig. 8).

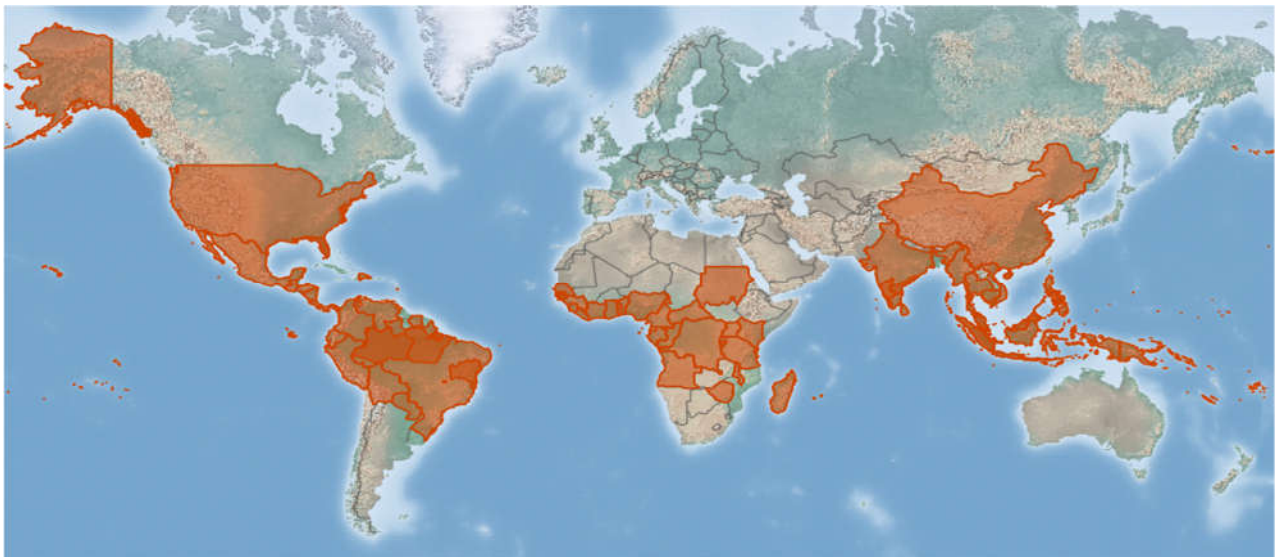


Fig. 8. Worldwide distribution of oil palm production (CABI, 2019).

In the last 60 years, the total oil production increased from 1.5 million tons in the 1960s to 75 million tons even more by 2020. Currently, the oil palm sector employs 6 million people directly and another 11 million indirectly worldwide, with an estimated yearly production value of US\$60 billion (Murphy *et al.*, 2020). According to estimates from various industry sources, the amount of palm oil needed by 2050 could range from 240–250 Mt (Babu *et al.*, 2021). Nowadays, the larger oil palm production is covered by both Indonesia and Malaysia, with a total production of 46.5 MT and 19.8 MT, respectively, which is almost 85% of the total world production. Thailand is the third-largest producer with a total production of 3.2 MT, sharing 4% of world production. In Africa, Nigeria, Ivory Coast, Ghana, and Cameroon are the larger producer of oil palm. Nigeria is the leading producer of oil palm with a total production of 1.4 MT, i.e., 1% of world production and followed by Cote d'Ivoire at 0.6 MT, Cameroon at 0.4

MT, and Ghana at 0.3 MT. The largest importers of palm oil are India, China, the European Union countries, and Pakistan (Statista, 2021).

I.2. BREEDING APPROACHES IN OIL PALM

Oil palm breeding can be undertaken by both conventional and modern biotechnological approaches, for instance, tissue culture, genetic modification, and MAS techniques (Corley & Tinker, 2016; Soh *et al.*, 2017). In any of the methods, the need for breeding is used to improve multiple traits affected by biotic and abiotic factors (Jalani *et al.*, 1997; Kwong *et al.*, 2016). The report from Ajambang *et al.* (2016) showed whatever the method used the main target of oil palm breeding is to improve the palm oil yield, reduction of vertical growth to extend the plant's economic life, select for tolerance to drought, select to diseases resistance (Fusarium wilt), and improve palm oil quality (high iodine value, low free fatty acids content). The conventional breeding method of oil palm has several limitation, for instance, costly, time-consuming, large areas required for planting, a long breeding cycle (around 20 years, while sexual maturity is reached at around 3 years of age), and a limited number of tested individuals (Wong & Bernardo, 2008; Cros *et al.*, 2014; Jin *et al.*, 2016; Seng *et al.*, 2016). In the same vein, Babu & Mathur (2016) showed that genetic improvement in oil palm through conventional breeding methods takes more than 12 years to obtain a new variety and lack of genetic homozygosity in current advanced parental breeding materials which results in many years to come up with new variety. Gain from breeding needs to be increased significantly to address new challenges such as land degradation, population growth, climate change, and agricultural land constraints as well as breeding cost. Several oil palm breeding options have been reported by different scholars, however, Corley & Tinker (2016); Florence *et al.* (2017); Soh *et al.* (2017); Ithnin & Din (2020) summarize mass selection, MRRS, tissue culture, genetic modification, and MAS are the major options of oil palm breeding. In the following paragraphs, the most practically used breeding options in the current era, mass selection, MRRS, and genomic selection are developed.

I.2.1. Mass selection

The genetic improvement through the application of a mass selection for oil palm for yield started in the 1920s in South-East Asia (Indonesia, and Malaysia), since then it is known as Belgian Congo (Mergeai, 2002; Corley & Tinker, 2016). It is the method of selection of best-

performed individuals based on their phenotypic value. So, the selection efficiency of individuals is mainly based on the heritability of targeted traits. A report from Soh *et al.* (2017) showed that this method started with a collection of open-pollinated seeds from the mass (forest) or phenotypically selected palms and followed by organized intercrosses among them, for example, West African (WA) $T \times T$ crosses and South East Asian (SEA) Deli $D \times D$ crosses. In the same vine, in SEA breeding through mass selection was established from one planting material four D seedlings were introduced into Java (Indonesia) in 1848 from an unknown part of Africa which resulted in a relatively homogenous and inbred breeding population called Deli. Furthermore, Deli is divided into several subpopulations, such as Marihat Baris, Elmina, etc (Mergeai, 2002; Durand-Gasselin *et al.*, 2011; Corley & Tinker, 2016). Similarly, Cochard *et al.* (2005) outlined that this SEA oil palm has, four to five generations, taking into account the generations of multiplication from individuals introduced in the Bogor Botanical Garden (Indonesia) in 1848.

In Africa, breeding through mass selection was less efficient compared to SEA, as it was complicated by the segregation of the fruit types in the crosses between the best breeding materials. In Africa, the sources of the breeding materials are D , T , and P types, which have been developed through the different breeding approaches used in SEA (Durand-Gasselin *et al.*, 2000; Corley & Tinker, 2016). According to Ajambang *et al.* (2016); Corley & Tinker (2016) reported that in Africa there are different breeding populations: La Mé (Côte d'Ivoire), Yangambi (Democratic Republic of Congo), Ekona (Cameroon), WAIFOR (Nigeria), etc. These populations were selected through mass selection, for instance, the La Mé population originated from 19 individuals selected from prosecutions made in the 1920s and the Yangambi population originated from 10 to 20 breeding materials in the 1920s, including the Djongo palm which given its exceptional qualities would have finally contributed more than 70% to the Yangambi population (Rivas *et al.*, 2012; Corley & Tinker, 2016).

In oil palm breeding through the aid of mass selection has resulted in a premounted result, for instance, a report from Corley & Tinker (2016) showed that some components of oil yield had a moderate level of narrow-sense heritability h^2 such as Mesocarp/Fruit (0.53) and BW (0.39) while other components (BN, F/B, and O/M) had low h^2 (< 0.25). In the same light, Soh (2012) depicted that breeding of oil palm through this method better resulted in the production of inter-population $D \times P$ (T hybrid) with the highly selected Deli D population having uniform high oil yielding and bigger bunches than the maternal parent and the highly selected WA (AVROS, Yangambi, La Mé, Ekona, NIFOR) P/T (uniform high oil yielding and high bunch number (BN)) as the paternal parent. According to Nyouma *et al.* (2019) reviewed

the century of breeding in oil palm and outlined that mass selection helps to get the current breeding populations grouped into two complementary groups (A and B) based on the characteristics of their bunch production. Group A, mostly from SEA (i.e., Deli *dura* population) and Angola, the latter is of lesser importance due to the production of 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. By the same token, Beirnaert & Vanderweyen (1941) outlined that mass selection breeding also helps to get a better understanding of the genetic control of the fruit type by a gene, nowadays called SHELL.

I.2.2. Modified reciprocal recurrent selection (MRRS)

Reciprocal Recurrent Selection (RRS) is fundamentally used as a population improvement method. It was first developed by Comstock *et al.* (1949) with the major goal to improve two different breeding populations of maize, for example, populations A and B, for combining well with each other and both populations subjected to selection at the same time. This approach mainly helps to improve heterosis breeding and not population improvement. The principle of this method mainly starts by selecting the base populations. The homozygous inbreeding lines of open-pollinated plants are obtained and conserved by continuous close selfing with the selection. Recurrent selection in the base populations would improve both the general and specific combining ability (SCA) of both populations. With continuous selection, the frequencies of desirable alleles and allele combinations that affect the trait of interest will increase. Moreover, the magnitudes of these allele increases will become larger with the increasing number of selection cycles (Dudley, 1997; Ithnin & Din, 2020).

In oil palm, the MRRS (Fig. 9) was first proposed by Gascon & De Berchoux (1964) by crossing A×B for bunch production, and its performance was more than 25% higher than the parental populations. Nowadays, MRRS has also been adopted and practiced by most breeding programs across the major oil palm cultivating countries namely the Nigerian Institute for Oil Palm Research (NIFOR), Ghana Oil Palm Research Institute (GOPRI), Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD) and coordinated programs established in Ivory Coast, Cameroon, Benin, Niger, and Guinea-Bissau. The method was also implemented in SouthEast Asia oil palm research centers, for example, the Indonesian Oil Palm Research Institute (IOPRI) and SOCFINDO in Indonesia based on breeding populations derived from CIRAD such as Deli, Angola, Yangambi, La Mé, and Yocobue. In

Malaysia, particularly in Applied Agriculture Resources (AAR) SdnBhd and MPOB. Even though it has various accomplishments among research centers, it generally follows the breeding scheme described in Fig. 9 (Soh *et al.*, 2017; Ithnin & Din, 2020).

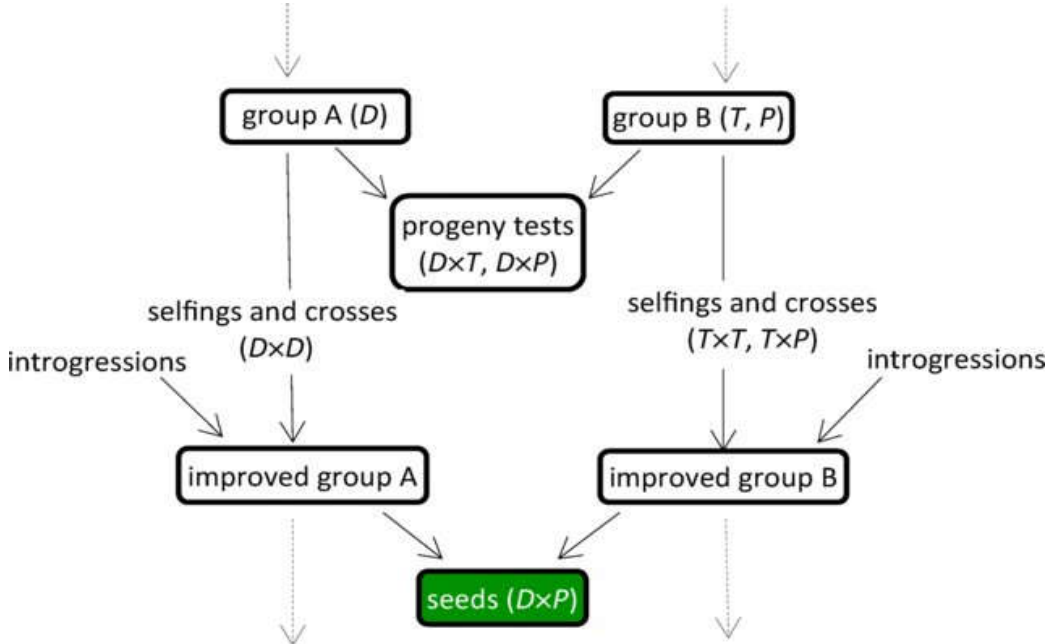


Fig. 9. Scheme of one cycle of modified reciprocal recurrent selection applied to oil palm. D: *dura*, T: *tenera*, P: *pisifera*, green: commercial seeds (Nyouma *et al.*, 2019).

According to Gallais & Poly (1990) the main advantages of recurrent selection are: increasing the frequency of genes and associations favoring the type of variety to be developed, enabling effective recombination, hence highly effective multi-trait breeding, preventing an over-rapid loss of variation, partially fixing heterosis, ensuring continuous, long-term progress, and providing outputs directly applicable for varietal creation. In one breeding cycle, compared to cereals (e.g. 3 months in rice) in oil palm application of RRS extends over a long period (~20 years) (Florence *et al.*, 2017). In oil palm, Durand-Gasselín *et al.* (2010) outlined that despite long breeding cycling time, a high genetic gain rate has been achieved since 1960 (~ +1%/year for yield). Similar to the other method, RRS has a limitation, for instance, Florence *et al.* (2017) pointed out the cost and time efficiency related to the estimation of the parental GCA. Conversely, Soh *et al.* (2017) due to severe inbreeding depression affecting seed and pollen production, RRS with each hybrid combination is usually limited to two cycles in oil palm. Despite the limitation, a review report Nyouma *et al.* (2019) and Soh *et al.* (2017) outlined that MRRS helps exploit the hybrid vigor for bunch production that appeared in the A × B crosses, and they enable better estimates of genetic values than mass selection. In general, Ithnin & Din

(2020), summarize the three major advantages of MRRS in oil palm, primarily it helps to select selfed parents for commercial hybrid seed production and further breeding activities i.e., commercial interpopulation hybrids parents development. Secondly, it helps to attain desirable alleles at the maximum level in the parents of both additive and non-additive. Thirdly, it helps the continuous commercial hybrid seed production due to the availability of data using *duras* and *pisiferas* that are generated from the self or sibs of the respective parents which are eventually planted simultaneously as the progeny tests.

I.2.3. Genomic selection

Nowadays, the availability of genetic and genomics resources generally in crop plants specifically in oil palm help to open a new door to apply new breeding tools called MAS (Rajinder & Choo, 2005; Collard & Mackill, 2008). By the same observation, Soh (2018) reviewed that the exposure of oil palm to whole-genome sequence helps the viability of MAS for many major QTL traits e.g. fruit color, mantling, long stalk, and lipase are being advanced. In MAS, molecular marker data can be used to predict phenotype(s), based on the known association between the chosen marker(s) and phenotype(s). Therefore, the marker-phenotype associations can be distinguished using a method called QTL mapping. However, for complex traits like oil yield which are governed by many large numbers of genes the efficiency of QTLs-based MAS is minimal, particularly for the plant having a small effective population size like oil palm (Muranty *et al.*, 2014). To this end, the genomic selection was developed as a specific case of MAS designed for quantitative traits (Meuwissen *et al.*, 2001).

Genomic selection (GS) is a MAS method with a high density of markers covering the entire genome so that at least one marker can be in LD with each QTL (Goddard & Hayes, 2007; Xu *et al.*, 2018). 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. Therefore, GS has emerged as one of the most promising selection strategies to enhance genetic gain, reduce breeding costs and time of breeding cycle for both animal and plant breeding programs, and it has several advantages as compared to both phenotypic and MAS (van der Werf, 2013; Voss-Fels *et al.*, 2019). Generally, GS is the most appropriate MAS method for yield traits which are usually quantitative, i.e. controlled by many loci with small effects. The establishment of a training set population is one of the first steps in a genomic selection which should consist of several hundred to a few thousand individuals that

are related to the validation population and with phenotypes for the traits of interest. The training population is genotyped for a genome-wide panel of markers and also phenotyped for the targeted traits, and a prediction model is developed using these genotypic and phenotypic data (Akdemir & Isidro-Sánchez, 2019). The selected population is also genotyped but not phenotyped, and the prediction model calculates the genomic estimated breeding values (GEBV) or genomic estimated genetic values of the selection population (Fig. 10). Therefore, GS has the potential to enhance genetic gain, increase gain per unit time, increase selection intensity, improve the accuracy, reduce breeding time and expenditure, reduce costs of genotyping, and enhance selection for low heritability traits (Cros *et al.*, 2015b; Grattapaglia & Resende, 2011; Resende *et al.*, 2017).

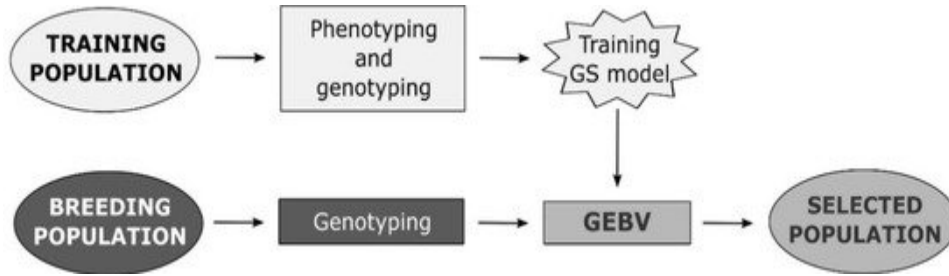


Fig. 10. Diagram of genomic selection (GS) processes (Plavšín *et al.*, 2021).

I.3. FACTORS AFFECTING THE ACCURACY OF GENOMIC SELECTION

The correlation between the genomic estimated breeding values (GEBV) and true breeding values (TBVs) is known as the GS accuracy of genomic (r_{GS}) (Equ.1) and it is an important parameter in GS due to its linear correlation with genetic gain R (Equ. 2) (Lin *et al.*, 2014):

$$r_{GS} = cor(TBVs, GEBVs) \quad [1]$$

$$Ry = \frac{i \times r \times \delta A}{y} \quad [2]$$

Where Ry is the annual genetic gain, i is selection intensity, r is selection accuracy, δA is the genetic standard deviation, and y is the generation interval in years.

The accuracy of GS is influenced by numerous factors Grattapaglia & Resende (2011); Isik (2014); Zhao *et al.* (2015); Nyouma *et al.* (2019); Robertsen *et al.* (2019): effective

population size, marker type and density, size and structure of the training population, the heritability of traits, genetic architecture, relatedness between training and validation population, LD between markers and QTLs, validation approaches and statistical model of prediction. In this section, we discuss how each of these factors affects the accuracy of GS in tropical perennial crops and plantation trees.

I.3.1. Linkage disequilibrium (LD) and effective size (N_e)

LD and N_e are the two interrelated effects that strongly influence GS accuracy (Heffner *et al.*, 2009; Isik, 2014; Lebedev *et al.*, 2020). LD is defined as the non-random association of alleles at two or more loci in haplotypes (Weir, 1979; Slatkin, 2008). LD between two loci is measured based on the frequency of alleles, using indexes like D , D' , and r^2 (Collins, 2007). A key assumption in GS is that there is LD between QTLs and markers, such that, with dense genome marker coverage, every QTL controlling the phenotype of interest would be in LD with at least one marker. Good knowledge of this parameter in the target population is therefore of particular interest to define the marker density required for GS. It is thus useful to explore historical events, such as bottlenecks, genetic drift, natural and artificial selection, that may have shaped the LD profile in the target population (Flint-Garcia *et al.*, 2003; Gupta *et al.*, 2005; Mackay & Powell, 2007; Slatkin, 2008). The LD profile is largely determined by the past N_e , which can be described as the number of randomly mating individuals in a population that would give rise to the observed rate of inbreeding (Falconer & Mackay, 1996). There is an inverse relationship between N_e and LD, with high rates of genetic drift and inbreeding in low N_e populations leading to strong LD between markers and QTLs compared to high N_e populations (Grattapaglia, 2014; Lin *et al.*, 2014; Thistlethwaite *et al.*, 2020). As N_e decreases and LD increases, pairs of individuals within the population tend to share longer haplotypes, enabling good genomic prediction accuracy (Heffner *et al.*, 2009; Clark *et al.*, 2012; Isik, 2014; Lebedev *et al.*, 2020). For a given marker density, training population size, and trait, LD and GS prediction accuracy is higher in populations with low N_e than in populations with high N_e (Solberg *et al.*, 2008; Grattapaglia, 2014; Lin *et al.*, 2014).

The crucial role of LD and N_e in GS accuracy has also been underlined in studies on tropical perennial crops and plantation trees. Several studies investigated the LD profile to evaluate whether the marker density was high enough in citrus Gois *et al.* (2016); Minamikawa *et al.* (2017), cocoa McElroy *et al.* (2018), eucalyptus Denis & Bouvet (2013); Durán *et al.* (2017); Müller *et al.* (2017) and oil palm (Kwong *et al.*, 2017a). Many studies in tropical perennial crops and plantation trees also investigated the efficiency of GS in populations with

high LD/low N_e . This was possible using populations obtained through specific mating designs among a reduced number of parents (Resende *et al.*, 2012; Denis & Bouvet, 2013). In this way, Resende *et al.* (2012) found that in a population of eucalyptus where $N_e = 11$ was obtained with an incomplete diallel, GS accuracy was higher for the four growth and wood quality traits studied than in the population where $N_e = 51$, despite a slightly larger number of training individuals in the latter population. In other studies, high LD/low N_e was obtained in full-sib families GS (Gois *et al.*, 2016; Cros *et al.*, 2017; Kwong *et al.*, 2017b; de Souza *et al.*, 2018). This strategy is also applied in other crops as it maximizes GS accuracy, although at the cost of only applying to families comprising the training population (Lin *et al.*, 2014; Crossa *et al.*, 2017; Lebedev *et al.*, 2020). The fact that GS accuracy reaches a plateau when marker density reaches a certain level (see below) suggests that an appropriate strategy to filter the markers would increase the cost-efficiency of GS. Filtering SNPs on LD has been investigated in several studies, as the SNPs that show very high LD values provide redundant information. In oil palm, Kwong *et al.* (2017a) evaluated the impact of marker density reduction by LD filtering and noted that, for some traits, it was possible to reach the same GS accuracy as using all the SNPs.

I.3.2. Marker density and type

As marker density strongly affects the extent of LD, it also plays a major role in GS accuracy. In GS studies of both plants and animals, increasing the number of markers was shown to improve prediction accuracy until a plateau was reached (Meuwissen *et al.*, 2001; Solberg *et al.*, 2008; Isik, 2014; Lin *et al.*, 2014; Robertsen *et al.*, 2019). The same trend was observed in tropical perennial crops and plantation trees, where the density of markers required to reach maximum prediction accuracy depends in particular on the type of population, trait, and marker. Romero Navarro *et al.* (2017) found increasing prediction accuracy for yield and disease traits in cocoa with increasing marker density before a plateau was reached at around 1,000 markers. In the rubber tree, the prediction accuracy for rubber yield plateaued at around 300 SSRs (Cros *et al.*, 2019). In eucalyptus, the prediction accuracy among five growth and wood property traits reached a plateau between 5,000 and 20,000 SNPs (Tan *et al.*, 2017). Among seven production traits in oil palm hybrids, the plateau was reached with 500 to 2,000 SNPs (Cros *et al.*, 2017).

GS accuracy is also affected by the type of marker. Thus, in oil palm, GS accuracy for BN and average bunch weight (BW) plateaued at 160 SSRs in heterotic group A and 90 SSRs in group B Marchal *et al.* (2016) versus 3,000 SNPs in group A and 350 SNPs in group B (Cros *et al.*, 2017). This likely resulted from the fact that, as SNPs are biallelic, they are less

informative than SSRs. However, in practice, SSRs cannot be used for genomic predictions, as SSRs rely on dense genotyping of large populations of selection candidates and therefore require high throughput genotyping approaches at a reasonable cost. If marker density is constrained by the genotyping approach, the GS accuracy may be reduced. Thus, Kwong *et al.* (2017b) obtained mean GS prediction accuracies of 0.21 over palm oil yield components using 135 SSRs, versus 0.31 with 200K SNPs. Two primary options are available to reach this goal with SNPs: methods that reduce genome complexity and SNP arrays. SNP arrays have been developed in several tropical perennial crops and plantation trees, with, for example, a 200K array in oil palm Kwong *et al.* (2016), a 60K array in eucalyptus Silva-Junior *et al.* (2015), and a 15K array in cacao (McElroy *et al.*, 2018). Most SNP genotyping methods based on reducing genome complexity consist of restriction enzyme-based approaches and sequence capture (Zhou & Holliday, 2012; Uitdewilligen *et al.*, 2013). These methods do not require specific preliminary investment and can be applied directly to any population, but are associated with a higher rate of missing data and genotyping errors than SNP arrays. Despite these differences, it seems that the choice between these two types of approaches has no impact on GS accuracy: the accuracy of genomic prediction of 13 wood quality and growth traits in eucalyptus using SNP genotypes obtained with sequence capture and a 60K SNP array was similar (de Moraes *et al.*, 2018).

I.3.3. Traits heritability

The broad-sense heritability of a trait (H^2) is defined as the proportion of the phenotypic variance that is genetically controlled. Narrow-sense heritability (h^2) considers only variations due to additive gene action and ignores non-additive (dominance and epistasis) genetic effects (Falconer & Mackay, 1996). In GS studies, the heritability of the trait affects the accuracy of GEBV, with higher h^2 leading to greater GS accuracy (Meuwissen *et al.*, 2001; Hayes *et al.*, 2009; Lin *et al.*, 2014). This was illustrated by studies in tropical perennial crops and plantation trees where positive correlations were found between h^2 and GS prediction accuracy for a set of disease resistance and yield traits in cacao Romero Navarro *et al.* (2017), eight palm oil production traits in the B heterotic group used in oil palm breeding Cros *et al.* (2015b), 18 Arabica coffee agronomic traits Sousa *et al.* (2019) and 15 vegetative growth, disease resistance, and fruit production traits in banana (Nyine *et al.*, 2018). When simulating GS in eucalyptus, Denis & Bouvet (2013) noted that the prediction accuracy was higher with $H^2=0.6$ than with $H^2=0.1$, regardless of the ratio of dominance to additive variance, modeling dominance, or not, or the breeding cycle. However, some studies detected no effect of trait

heritability on GS prediction accuracy, but the effect may have been masked by other factors with stronger effects on prediction accuracy than heritability, in particular variations in the size of the training population, among traits, like in (Durán *et al.*, 2017).

I.3.4. Statistical models for genomic prediction and trait genetic architecture

The whole-genome regression models used for genomic predictions deal with the ‘*large p, small n*’ problem, that, in GS, concerns the number of markers that usually (largely) exceeds the number of data records, in contrast to multiple linear regressions that cannot be used without variable selection, which conflicts with the original goal of GS, i.e. avoiding marker selection and overfitting. Multiple linear regression results in an insufficient degree of freedom leading to poor prediction due to the inability to estimate all marker effects at the same time, which is exacerbated by multicollinearity. A wide range of statistical methods has been developed for GS to alleviate this constraint (Jannink *et al.*, 2010; Campos *et al.*, 2013; Morota & Gianola, 2014; Wang *et al.*, 2018; Montesinos-López *et al.*, 2021; Tong & Nikoloski, 2021). They represent two broad categories: (i) parametric approaches, which mainly include methods that rely on the best linear unbiased prediction methodology (genomic BLUP [GBLUP] and random regression BLUP [RRBLUP]) and various Bayesian methods (Bayesian LASSO, BayesA, BayesB, etc.), and (ii) semi- and non-parametric approaches, that fall into the machine learning category (reproducing kernel Hilbert spaces [RKHS], artificial neural networks, etc.). These methods differ in several ways: in terms of genetic assumptions and modeling of the genetic architecture of the traits (e.g., purely additive models, models that explicitly model dominance and/or epistatic effects, models with marker effects sampled from a common statistical distribution [RRBLUP, GBLUP], models with marker effects sampled from specific distributions [Bayesian LASSO, BayesB, etc.], models that implicitly model non-additive effects [e.g. RKHS]), in terms of computational approach (relationship-based methods and marker effect-based methods, single trait and multi-trait models, etc.), and terms of the genomic information used in the model (type of polymorphisms, use of a priori information on markers, a combination of omics data, etc.). The most widely used statistical approach for GS is GBLUP Heslot *et al.* (2015); Montesinos-López *et al.* (2021), which combines linear mixed model analysis and genomic relationships.

The relative performance of the different statistical methods is expected to vary depending on the genetic architecture of the trait considered (Lebedev *et al.*, 2020). Genetic architecture corresponds to the genetic characteristics that determine the genotype-phenotype

relationship, in particular, the number of genes that control the trait, the number of alleles per gene, the distribution of the genes along the genome, the distribution of the gene effects, and the mode of gene action (additive, dominant, epistatic) (Momen *et al.*, 2018). Thus, methods in which marker effects are sampled in distributions where variance is the same for all markers (e.g. GBLUP, RRBLUP, Bayesian random regression) are expected to be more suitable for traits following the infinitesimal model, while methods with marker-specific variances (e.g. Bayesian LASSO, BayesB) are expected to be more suitable for traits whose genetic architecture includes major QTLs. Consequently, many GS studies, including those on tropical perennial fruit crops and plantation trees, use a range of statistical prediction methods to identify the most appropriate one for a specific trait. Overall, few variations have been found among statistical approaches, for example, in oil palm yield components Cros *et al.* (2015b); Kwong *et al.* (2017a), in eucalyptus growth Durán *et al.* (2017); Müller *et al.* (2017) and rubber tree latex yield (Cros *et al.*, 2019). This confirms results obtained in empirical evaluations in other species, in which GS statistical methods were seen to perform similarly Heslot *et al.* (2015), however, in some cases, differences were found: e.g., BayesB performed best for several traits including vegetative growth, production, and disease resistance in banana Nyine *et al.* (2018) and vegetative growth and oil yield in oil palm (Ithnin *et al.*, 2017). This could mean that in the populations considered, QTLs with large effects were segregated for these traits.

Similarly, when non-additive effects play a significant role in genetic variation, models that account for non-additive effects are expected to increase GS accuracy. In a simulation study, Denis & Bouvet (2013) showed that modeling dominance for the genomic predictions of the genetic value of eucalyptus clones improved accuracy when dominance effects were preeminent (ratio of dominance to the additive variance of 1.0) and heritability was high ($H^2 = 0.60$). With empirical data, also in eucalyptus, Resende *et al.* (2017); Tan *et al.* (2018); Paludeto *et al.* (2021) showed that the use of GS models that account for dominance increased the accuracy of prediction for growth traits, which had high levels of dominance variance, whereas this was not the case for wood traits. In citrus, Minamikawa *et al.* (2017) showed that considering both additive and dominance effects improved prediction accuracy for acidity and juiciness.

When considering traits correlated with a sufficient magnitude but with contrasting levels of heritability, the use of multi-trait models can increase prediction accuracy for low heritability traits (Tong & Nikoloski, 2021). In tropical perennial crops and plantation trees, the results obtained in oil palm Marchal *et al.* (2016) and Eucalyptus robusta Rambolarimanana *et al.* (2018) agreed with this principle. Multivariate models thus offer the opportunity to improve

prediction accuracy at no extra cost (apart from increased computational resources), and they should therefore be systematically evaluated when correlations exist among the traits of interest, or between the traits of interest and secondary traits.

Machine learning methods are complex black-box approaches that are of growing interest for genomic predictions as they have several desirable features. They avoid the use of assumptions that are often violated and cannot be verified Gianola & Van Kaam (2008) and they are particularly suitable to account for non-additive effects in particular in polyploids Bayer *et al.* (2021) and to integrate data from different biological sources for multi-omics predictions (Montesinos-López *et al.*, 2021; Tong & Nikoloski, 2021). RKHS is the most often evaluated machine learning approach for GS in tropical perennial crops and plantation trees. In bananas, RKHS was slightly more accurate than parametric approaches for a few traits (Nyine *et al.*, 2018). In a study analyzing eight traits in *E. urophylla* × *E. grandis* eucalyptus hybrids, RKHS proved to be slightly more accurate in predicting low-heritability traits but less accurately in predicting pulp yield Tan *et al.* (2017) and performed similarly to GBLUP for three traits in *E. grandis* (Rambolarimanana *et al.*, 2018). A few other machine learning methods have been implemented in tropical perennial crops and plantation trees. Maldonado *et al.* (2020) compared several parametric prediction models, RKHS and two artificial neural network approaches, deep learning, and Bayesian regularized neural networks, in *E. globulus* and maize, and found that predictions made with deep learning methods were significantly more accurate for all the traits considered. Sousa *et al.* (2020) compared several machine learning approaches and a parametric model to predict resistance to leaf rust in *Coffea arabica* and obtained the best accuracy with artificial neural networks. Several authors used random forest in oil palm and citrus and found that on average over several traits, random forest performed no better than parametric approaches (Kwong *et al.*, 2017b; Minamikawa *et al.*, 2017). In oil palm, the support vector machine was found to be slightly better on average than other methods (Kwong *et al.*, 2017b). Despite these uneven results in tropical perennial crops and plantation trees, machine learning should be further investigated, in particular as the training populations used so far was possibly not large enough for the optimal training of this type of approach (Montesinos-López *et al.*, 2021). Particular attention should also be paid to artificial neural networks, which have produced promising results. One limit to the differences among statistical methods and models in perennial fruit and tree crops reported so far is that they were not always supported by a statistical test indicating whether the differences were significant or not. This can be done for example using the Hotelling-Williams t-test (Steiger, 1980).

I.3.5. Training and validation population relatedness

The accuracy of GS is positively correlated with the relatedness between the training and test population (Pszczola *et al.*, 2012; Daetwyler *et al.*, 2013; Wientjes *et al.*, 2013; Isidro y Sánchez & Akdemir, 2021). This is because when pairs of genotypes are closely related, they tend to share long haplotype blocks in the same linkage phase. To limit allele duplication and redundancy, relationships within the training population should be minimized (Isidro y Sánchez & Akdemir, 2021). The accuracy of GS in tropical perennial crops and plantation trees was also found to be affected by the relatedness between the training and test population. In two eucalyptus species, *E. benthamii* and *E. pellita*, Müller *et al.* (2017) found that prediction accuracy declined strongly for three growth traits when individuals were randomly assigned to the training and validation populations compared to when they were assigned using a principal component analysis to minimize relatedness between training and validation populations. Similarly, considering eight wood growth and quality traits in *Eucalyptus urophylla* × *E. grandis*, Tan *et al.* (2017) obtained the worst prediction accuracies when minimizing the relatedness between the training and validation populations using k-means clustering. In another study, a significant positive correlation was found between GS accuracy and the relationship between training and validation populations for various production traits in oil palm (Cros *et al.*, 2015b).

I.3.6. Size and design of the training population

The size of the training population is one of the most important factors that determine GS accuracy. Several GS studies have reported that increasing the size of the training population improves GS accuracy (Combs & Bernardo, 2013; Isidro *et al.*, 2015; Nielsen *et al.*, 2016; Tan *et al.*, 2017; Cericola *et al.*, 2018; Calleja-Rodriguez *et al.*, 2020). In a family of full-sibs of *Hevea brasiliensis*, Cros *et al.* (2019) reported an increase in the accuracy of GS for rubber yield with an increase in the size of the training population up to a plateau of 200 individuals. In *Eucalyptus*, Denis & Bouvet (2013) also reported an increase in GS accuracy as a result of increasing the size of the training population, and Tan *et al.* (2017) reported an increase in GS accuracy that followed a diminishing return trend with increasing size of the training population.

The possibility of assembling large training populations among tropical perennial crops and plantation trees is contrasted. Thus, training populations comprising more than 1,000 individuals were used in eucalyptus Mphahlele *et al.* (2021), cacao McElroy *et al.* (2018), and oil palm Kwong *et al.* (2017a), whereas only small populations (< 600 individuals) have been

used so far in banana Nyine *et al.* (2018), rubber tree Cros *et al.* (2019); Souza *et al.* (2019); Munyengwa *et al.* (2021), coffee Ferrão *et al.* (2019); Sousa *et al.* (2019), p. 2; Fanelli Carvalho *et al.* (2020); Sousa *et al.* (2020), jatropha Peixoto *et al.* (2017) and guava (Silva *et al.*, 2021). However, the size of the training population must be considered with the relatedness between training and validation populations. Thus, for GS predictions in a biparental cross, it is better to use a relatively small but highly related training population of full-sibs or half-sibs than a large training population comprising distantly related or unrelated individuals (Brandariz & Bernardo, 2019; Brauner *et al.*, 2020).

For some of the species considered here, breeding relies on a large number of phenotyped individuals, e.g., thousands of individuals for yield components and tolerance to ganoderma disease in oil palm Cros *et al.* (2017); Daval *et al.* (2021), and thousands of individuals for tolerance to pests and diseases in *Eucalyptus grandis* (Mphahlele *et al.*, 2021). In this case, genotyping a sample of the phenotyped population and making the genomic predictions using the single-step GBLUP approach Lourenco *et al.* (2020), i.e. using a training population combining the genomic data of the genotyped individuals and the genealogical data of the others, is an efficient way to maximize the cost efficiency of GS, see Mphahlele *et al.* (2021) in *E. grandis*, Cappa *et al.* (2019) in a complex eucalyptus population, and Imai *et al.* (2019) in citrus. The cost of phenotyping is a major constraint in GS, especially now that sequencing costs have dramatically decreased thanks to next-generation sequencing (Akdemir & Isidro-Sánchez, 2019). This financial constraint is particularly applicable to perennial crops, as their phenotypic evaluation requires large surface areas over several years. Thus, training populations need to be optimized to improve the cost-effectiveness of GS in these species. Training population optimization is the process of selecting, within a pool of individuals that could be used to train the GS model, a sample of individuals that will best predict the genetic value of the selection candidates (Isidro y Sánchez & Akdemir, 2021). Several methods have been developed to optimize the training population, including CD-mean, PEV-mean, stratified sampling, or EthAcc (Isidro y Sánchez & Akdemir, 2021). This aspect has received little attention in tropical perennial crops and plantation trees, although in oil palm, Cros *et al.* (2015b) confirmed the efficiency of training population optimization to improve GS accuracy.

I.4. BASIC CONCEPT OF POPULATION GENETICS

Population genetics is a sub-area of biology that investigates the genetic makeup of biological populations and how that makeup varies as a result of various causes, such as natural

selection (Maia & de Araújo Campos, 2019). It is the study of genetic variation within and between populations, as well as the evolutionary causes that contribute to this diversity. It is based on the Hardy-Weinberg law, which is true as long as the population size is high, mating is random, and mutation, selection, and migration are minimal (Johnston *et al.*, 2019). As a result, population geneticists work toward their goals by developing abstract mathematical models of gene frequency dynamics, attempting to derive inferences about the patterns of genetic variation in real populations from those models, and then correlating their findings with empirical evidence. Several parameters can be conserved to access the genetic constitution of a population, such as phenotypic frequencies, genotype frequencies, allelic frequencies, gene flow, heritability, genetic correlation, genetic diversity, and heterozygosity, which allow an understanding of the population's genetic dynamics (Maia & de Araújo Campos, 2019). Even though the concept of population genetics is broad and has numerous factors that alter the genetic makeup of the entire population, for instance, mutation, migration (with gene flow), natural selection, genetic drift, etc Kimura (1983), for this specific section, I try to explore the part and parcel of population genetics concepts like LD, Ne, haplotype sharing, and F_{st} .

I.4.1. Linkage disequilibrium

LD Lewontin & Kojima (1960), is defined as the nonrandom association of alleles at two or more loci in haplotypes (Bernardo, 2010). It supplies information about the history of the population associated with both natural and artificial selection. LD throughout the genome provides information about population history, the breeding system, and the pattern of geographic subdivision, while LD in the specific genome region gives information about the history of natural selection, gene conversion, mutation, and other forces that cause gene-frequency evolution (Slatkin, 2008; Goode, 2011). Consider the two linked loci Locus 1 has alleles A_1 and A_2 occurring at frequencies p_1 and p_2 and the same Locus 2 in another haplotype with alleles of B_1 and B_2 occurring at frequencies q_1 and q_2 in the population. Therefore, the possible haplotype combination can denote as A_1B_1 and A_1B_2 with frequencies h_{11} and h_{12} . In this regard, the two loci are linked together and produced a new haplotype by the process called linkage equilibrium (LE). However, if the occurrence of alleles A in the i^{th} alleles and B in the j^{th} alleles the haplotype passes independently, i.e does not pass randomly, and the two alleles are produced a new haplotype by the process called LD (Calabrese, 2019). LD is the intensity between two loci and is measured based on the frequency of alleles by using indexes like D, D' and r^2 and it ranges from completely random ($|D|=|D'|=r^2=0$) to complete LD ($|D|=0.25$, $|D'|=r^2=1$), r^2 can range from 0 (two loci in equilibrium) to 1 (non-random loci in complete LD)

(Equ. 3) (Collins, 2007; Nakaya & Isobe, 2012). The result of LD can be positive or negative values, if LD is positive it indicates the two alleles occur together on the same haplotype and negative when the two alleles occur together on a different haplotype (Calabrese, 2019). The LD measure for D , D' and r^2 for a biallelic locus with alleles A and a , at locus 1; B and b at locus 2, is explained in the following formulas as follows (Slatkin, 2008):

$$D' = P(AB) - P(A)P(B) \text{ observed} - \text{predicted}$$

$$D = P(AB) * P(ab) - P(Ab)P(aB)$$

[3]

$$r^2 = \frac{D^2}{P(A) * P(a) * P(B) * P(b)}$$

Where, there are two loci each with two alleles (A , a , and B , b) and $P(AB)$ is the frequency of the AB haplotype.

In GS, the extent of LD between markers and QTL is important and it is one of the major factors affecting the accuracy of GS and a good knowledge of this parameter helps to define the marker density required for GS (Heffner *et al.*, 2009; Isik, 2014). Indeed, the concept of GS relies heavily on LD between QTLs and DNA markers and it is thus useful to explore the potentially significant historical events that occurred during domestication, breed formation, and natural and ongoing selection in the target population (Li & Kim, 2015; Jemaa *et al.*, 2019). High-resolution LD maps are important to provide useful information for high-density SNP design panels in GS (Bejarano *et al.*, 2018). Generally, strong LD results in higher prediction accuracy (Wientjes *et al.*, 2013). Numerous factors are affecting LD, for instance, the rate and type of inbreeding in a given species, the size of the population that we analyzed, genetic drift, mutation rate, recombination frequency, the extent of population stratification, and subdivision (Qian *et al.*, 2017).

I.4.2. Effective population size

Effective population size is the number of randomly mating individuals giving rise to the observed rate of inbreeding in a given population (Falconer & Mackay, 1996). It is also defined as the number of individuals who are actively involved in generating the following generation (Sbordoni *et al.*, 2004). It is a measurement of the number of independent breeding individuals in a population (Corbin *et al.*, 2012). A lower N_e results in higher rates of inbreeding in a population which ultimately leads to genetic drift (Poets *et al.*, 2015). Equally important,

Lin *et al.* (2014) reviewed that the lower N_e of a given population is subjected to strong genetic drift, which results in one of the major factors influencing LD between loci. N_e determines the accuracy of GS through its effect on the extent of genome-wide LD (Meuwissen *et al.*, 2001; Hayes & Goddard, 2010). There is an inverse relationship between LD and N_e . In populations with lower N_e , LD is high due to higher genetic drift (Lin *et al.*, 2014). On the contrary, Goddard *et al.* (2010) pointed out in populations with large N_e result in a lower LD and GS accuracy, because more markers are required to establish linkages between marker and QTL to sustain the power of predication by marker across breeding populations and the growing environment. Generally, keeping the other factors constant, as the N_e is reduced, leading to an increase in the extent of LD between markers and QTL gives rise to better accuracies. Wright (1931) produced a general equation to calculate the effective population size in a population with nonoverlapping generations and unequal sex ratios:

$$N_e = \frac{4N_m N_f}{(N_m + N_f)} \quad [4]$$

where N_m and N_f are the numbers of males and females, respectively, contributing to the *gamete* pool each generation

N_e has a strong impact on tree breeding GS and determines the number of markers needed to optimize accuracy (Denis & Bouvet, 2013). Likewise, N_e plays a pivotal role in GS study in tropical perennial fruit and forest tree crops. A review report Isik (2014), showed that breeding programs in a tree are conservative, and maintain >200 individuals (status number), with the main goal of genetic conservation. Furthermore, White *et al.* (2007), in forest tree breeding N_e of the base population is generally high and declines upon successive breeding cycles by applying strong selection pressure. By the same token, a report from Namkoong *et al.* (2012) showed that N_e between 20 and 50 in tree breeding populations helps to get better selection intensity for a desirable trait with significant genetic gains for several generations. So far, in oil palm, N_e was only estimated in the Deli population Cros *et al.* (2014) and there is no information about N_e for the La Mé and other populations.

I.4.3. Haplotype sharing

The term haplotypes are defined as “two or more SNP alleles that tend to be inherited as a unit in the chromosome” (Gabriel, 2002; Bernardo, 2010). Lloyd *et al.* (2016) also defined the term haplotype sharing as any combination of alleles or markers, such as SNPs, without regard to their reproducibility, inheritance, polymorphism, or biological significance. In other terms, a haplotype is a collection of neighboring genomic structural changes, such as

polymorphic SNPs, that are in substantial LD (Maldonado *et al.*, 2019). It helps to reproduce the genetic resemblance between individuals and it is a natural extension of identity by descent, a measure of genetic resemblance for individuals in a pedigree, to unrelated samples (Xu & Guan, 2014). For a given stretch of chromosomal DNA, each individual has two haplotypes; yet, at the population level, numerous haplotypes might be detected for the same stretch (Bhat *et al.*, 2021). As described in Fig. 11 a haplotype is made up of two or more polymorphic SNPs from haploid sequences that are inherited as a unit.

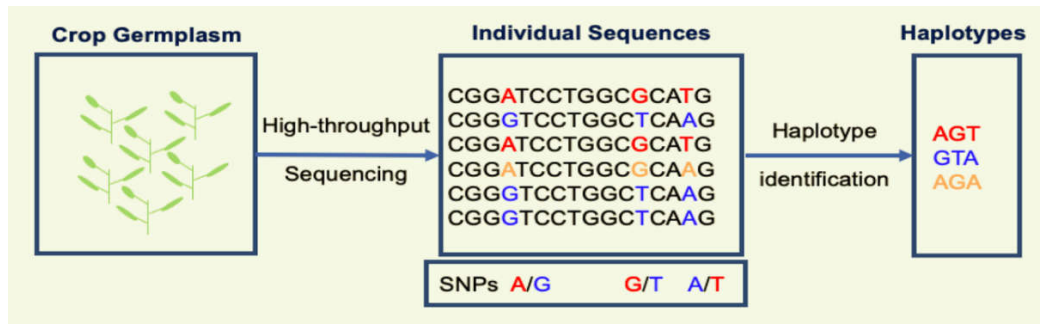


Fig. 11. Formation and development of haplotypes from haploid sequences (Bhat *et al.*, 2021).

Haplotypes-based breeding has the potential for whole-genome selection and helps to identify genomic regions related to the trait of interest in a population with their specified LD structures (Varshney *et al.*, 2005). Qian *et al.* (2017) outlined in their review paper that the identification of breeding lines with their haplotype block helps to separate favorable and unfavorable genetic variation. Haplotype-based crop improvement also helps to identify the genomic region between the current commercial and the elite breeding genetic materials in crop species (Bevan *et al.*, 2017). On the other hand, haplotype-based genomic selection also helps to improve the accuracy of genomic selection (Qian *et al.*, 2017). Hamazaki & Iwata (2020) haplotype-based GS improves the accuracy of the GS due to the haplotypes can better capture LD and genomic similarities in various lines and may catch local high-order allelic interactions. Similar results were reported by other researchers on different crops, for instance, wheat Sallam *et al.* (2020), and sorghum (Jensen *et al.*, 2020). Generally, Bhat *et al.* (2021) summarized that haplotype-based GS helps to increase the accuracy of GS and it is important to encourage GS researchers to focus on crop improvement programs with haplotype-assisted genomic prediction. So far there are no haplotype research findings in oil palm.

I.4.4. Fixation index (F_{st})

F_{st} is used to identify loci with divergent allelic frequencies between two or more populations (Wright, 1978). It is the genetic differentiation coefficient that ranges from 0, (no variation between populations and/or signifying a deficiency of heterozygotes) to 1 (each population is fixed with a different allele and/or indicates an excess of heterozygotes) (Hayati *et al.*, 2004). It helps to understand the genetic differentiation among groups of a population (Jakobsson *et al.*, 2013). It is the applicable index to study gene-frequency differentiation if, the genetic diversity is low (Nagylaki, 1998). The F_{st} analysis can improve GS and GWAS studies. For example, Chang *et al.* (2019) showed that prioritizing and weighting SNPs based on their F_{st} values can increase the accuracy of genomic predictions by more than 5%. Yan *et al.* (2017) in soybean reported that selective genotyping genome-wide association study (GWAS) and F_{st} analysis helped to identify QTLs for seed weight. Wright (1978) F_{st} has been used to measure the level of genetic differentiation between populations based on the change in allele frequencies, and the general formula for F_{st} is designated in the following equation as follows:

$$F_{ST} = \frac{\text{Variation between populations} - \text{Variation within populations}}{\text{Variation between populations}}$$

$$F_{ST} = \frac{H_T - H_S}{H_T} \quad [5]$$

Where H_S is the average heterozygosity in the subpopulations, H_T is the average heterozygosity in the metapopulations.

I.5. OIL PALM GENOME MAPPING

Genetic linkage maps reflect the actual inheritance of loci from parents to their offspring based on the patterns of recombination during meiosis. In oil palm for the last 20 years, several linkage maps have been constructed using different marker densities and types, population structures, and software and used to detect different vegetative, yield, and yield components and palm oil quality traits (Ong *et al.*, 2019). In the same light, markers like RFLPs, AFLPs, SSRs, and SNPs are widely used to construct genetic linkage maps in oil palm, and very recently restriction-associated DNA tagging (RAD), double digestion RAD (ddRAD), single primer enrichment technology (SPET) has been recognized for producing a large number of SNPs with

remarkable maps Herrero *et al.* (2020) shows an outline of the major studies on oil palm genome mapping with their different features.

In oil palm, the first genetic linkage map constructed based on Restriction Fragment Length Polymorphism (RFLP) markers from genomic libraries was published in 1997 (Mayes *et al.*, 1997). This map which considers 97 RFLP markers (84 probes) mapped a selfed *guineensis* cross (*tenera* x *tenera*) with a total genetic distance of 860 cM producing a total of 24 linkage groups (LGs) using a logarithm of the odds (LOD) score of 4 and recombination fraction of 0.4. According to the study Mayes *et al.* (1997), more than 95% of the markers could be linked to at least one other marker, suggesting that good genome coverage helps to detect the position of the shell thickness gene (*Sh*) at a distance of 9.8cM on group 10. From their result, Mayes *et al.* (1997) concluded that this map helps to enable the mapping of the gene responsible for controlling major commercial oil palm traits. Likewise, Rance *et al.* (2001) also used 153 RFLPs markers to construct a genetic linkage map of 84 self-fertilization F₂ oil palm populations used to detect major genes influencing shell thickness. This map produced a total of 22 LGs giving a total map length of 852 cM using a LOD score of 4 and a recombination fraction of 0.49. The result confirms that QTL mapping helps to detect genes that influence a large proportion of the total phenotypic variance in a large and small population.

Further, Random Amplified Polymorphic DNA (RAPD) is another marker that is used to construct a genetic linkage map in oil palms. The first RAPD marker map was developed by Moretzsohn *et al.* (2000) to develop a pseudo-testcross mapping strategy in combination with the RAPD assay to construct genetic linkage maps of different fruit types (shell thickness) of F₁ *tenera* (*Sh*⁺*Sh*⁻) x *pisifera* (*Sh*⁻*Sh*⁻) progeny populations. The map used a total of 48 RAPD markers, and 308 F₁ progeny populations, and produced a total of 12 LGs with a map distance ranging from 399.7- 449.3 cM at a LOD score of 5.0 and by considering the projected *Elaeis* total map distances and genome sizes, physical and genetic distances relationships were established (1.06 Mbp/1 cM and 1.09 Mbp/1 cM, for *tenera* and *pisifera*, respectively). They also obtained limited genome coverage with the two maps (28.0%, for *tenera* and 25.6%, for *pisifera*). This result depicted the importance of RAPD markers used for genetic linkage mapping markers closer to the *Sh*⁺ locus, helped to detect the gene responsible for shell thickness, and gave a step forward for MAS for shell thickness in the oil palm.

Amplified fragment length polymorphism (AFLP) is another pronounced marker used to construct a genetic linkage map in the oil palm. The first AFLP based genetic map in oil palm was developed by Billotte *et al.* (2005) involving a cross between a thin-shelled *E. guineensis* (*tenera*) palm and a thick-shelled *E. guineensis* (*dura*) palm with the main goal of

mapping to detect the presence and absence of gene responsible for shell in the oil palm fruit. For this purpose, they used a total of 944 markers (255 SSRs, 688 AFLPs, allele *sh*) markers with a map length of 1,743cM and with an average of one marker every 1.8cM and LOD score of 3.0, producing a total of 16 LGs. The lengths of the LGs varied between 59 cM and 192 cM. Based on their finding, the application of a high-density linkage map is used to step forward research for QTLs and physical mapping in the *E. guineensis* species. This map was the first linkage map for the oil palm to have 16 independent LGs corresponding to the haploid chromosome number of 16 in the oil palm. Besides, they also reported that SSRs markers had better mapping resolution compared to that of AFLPs. This is because high-density markers like SSRs have higher recombination rates than low-density markers like AFLPs. From the result, they observed that SSRs markers are more well distributed along the genome than AFLPs markers. Conversely, Singh *et al.* (2009) reported an interspecific cross involving Colombian *Elaeis oleifera* (UP1026) and a Nigerian *E. guineensis* (T128), and a total of 118 palms from this interspecific cross were used to detect quantitative trait loci (QTLs) controlling oil quality (measured in terms of iodine value and fatty acid composition). To analyze the map, they used a total of 252 markers (199 AFLP, 38 RFLP, and 15 SSR) with a map length of 1815cM and with an average interval of 7 cM between adjacent markers, producing a total of 21 LGs with an average number of 12 markers per LGs. Again, almost in all maps, the markers were distributed at an interval of 25 cM except for linkage group 17 having 30 cM, indicating that the map is relatively homogeneous with regards to marker distribution; this is useful for tagging traits of economic interest for MAS. In this map, the length of individual LGs varied from 26.1 cM to 168 cM, with an average of 94cM. The application of the genetic linkage map helps to detect QTLs for fatty acid composition in oil palm and serves as a tool for the MAS breeding program.

Similarly, a report from Seng *et al.* (2011) used a total of 120 hybrid crosses between high-yielding *dura* (ARK86D) x *pisifera* (ML161P) using AFLP markers. To construct the map, they used a total of 479 marker loci (331 SSRs, 142 AFLPs, and 6 PCR–RFLPs) and 168 anchor points with a map length of 2,247.5 cM and an average map density of 4.7 cM using a LOD score of 3.0. They constructed a total of 16 LGs from 15-57 markers per linkage group with an average of 29 markers per linkage group and with lengths ranging from 77.5 cM to 223.7 cM, and an average of 137 cM. In line with this, the markers were well distributed all over the 16 LGs. Out of these, only LGs 3 and 9 have a long interval compared with others with 26.9 and 25.6 cM lengths, respectively. From their findings, Seng *et al.* (2011) concluded that the application of a genomic map in oil palm helps to validate against a closely related

population and helps identify yield-related QTLs. Likewise, Ting *et al.* (2013) and Ukoskit *et al.* (2014) also used the AFLP markers to construct a genetic linkage map in oil palms.

Simple sequence repeat (SSR) markers are co-dominant molecular markers that distinguish polymorphism and mapping in the oil palm genome. In the year 2005, SSR markers were used for the first time to construct a map of the oil palm. To construct the map, Billotte *et al.* (2005) used a total of 255 SSR markers with a map length of 1,743 cM with an average marker density of 7 cM. using a LOD score of 3.0 and producing a total of 16 LGs. Based on the outcome of their finding, mapping of oil palm using high-density markers like SSR brings milestone information for QTL mapping and other MAS research in the oil palm. In line with this, Billotte *et al.* (2010) used an SSR marker for QTL detection with a multi-parent linkage map of the cross (within-family analysis and across-family analysis) between two oil palm populations. They used a total of 150 palms in the controlled cross between Africa (LM2T) x Deli (DA10D). To construct the map, a total of 251 SSR markers were used. Based on their finding, the SSR map for LM2T x DA10D had 16 LGs and 253 loci, with a map length of 1,479 cM. and an average marker density of 6 cM. The large mapping genome was found in LG4 with spanned 134 cM on an average range of 61-250 cM and around 47% of the mapped loci had three or four alleles with an average density of 32 cM on the genome. In conclusion, a total of 156 SSRs (45 %) and the *Sh* locus were mapped and the mapping of the crossed oil palm populations helped to identify the QTL locus for the major gene-controlling fruit shell (*Sh*).

By the same token, Montoya *et al.* (2013) used a total of 347 segregating SSRs, 14 SNPs of genes, and the *Sh* locus to establish the linkage map and to detect QTLs of palm oil fatty acid composition. They produced a total of 16 LGs with a relative map length of 1485 cM and an average marker density of 4 cM at LOD 7.5 with a maximum recombination threshold of 0.3. Depending on their position in the linkage group, the length of the LGs ranged from 49.1 to 175.9 cM, with an average of 92.8 cM. Concerning QTLs, a total of 19 QTL associated with the palm oil fatty acid composition was obtained and this mapping helped to identify key genes in the oil palm genome related to oleic acid C18:1. In conclusion, 73 % (253) of the mapped SSRs segregated only from the hybrid parent SA65T, 2 %, (7) from PO3228D only, and 27 % (93) were common SSRs segregating from both parents. Again, the high number of mapped SSR loci with accurate relative linear orders, and their molecular hyper-variability helped to undertake other such mapping studies in other *Elaeis* breeding materials.

Later on, Cochard *et al.* (2015) constructed a linkage map using a 281 SSRs marker and a total of 271 genotyped oil palm populations. They produced a total of 16 LGs covering group A (2078 cM) and group B (1845 cM), with an average density of one marker every 9 and 7 cM,

respectively. Generally, the integrated maps gave a total map length of 1935 cM with a total of 281 markers and an average density of one marker for every 7.4 cM. Besides, the marker orders between physical and genetic maps were in good accordance, except for some sporadic markers. Based upon their finding they concluded that this output could help to step towards efficient pedigree-based QTL mapping using the first intercrossed generations in current breeding programs. Similar studies have been done using SSRs for the mapping of the oil palm genome. For instance, QTLs identification is associated with callogenesis and embryogenesis Ting *et al.* (2013), QTLs mapping for oil yield using African oil palm Jeennor & Volkaert (2014), linkage map, and QTLs analysis for sex ratio and related traits Ukoskit *et al.* (2014), genetic maps construction for two independent oil palm hybrids Ting *et al.* (2014), linkage mapping and identification of major QTL genes for stem height Lee *et al.* (2015) which all brought remarkable results for the oil palm genome mapping and molecular breeding research.

Currently, in oil palm single nucleotide polymorphisms (SNPs) are the most highly preferred and high-density markers used to study genetic diversity and population structure, construct high-density genetic maps, and provide genotypes for the genome-wide association Xia *et al.* (2019), and genomic selection studies (Cros *et al.*, 2018, 2017; Nyouma *et al.*, 2020). The first SNPs marker-based oil palm genome mapping was constructed by Jeennor & Volkaert (2014) using a total of 190 segregating loci (89 SSRs, 90 genes, and 11 non gene-based SNP markers), which were mapped into 31 LGs by applying threshold LOD of 3 and a recombination fraction of 0.45. They produced a map with a total length of 1,233 cM containing two to 20 markers covering a length between 1.5 and 103.5 cM, and with an average distance between markers of 6.5 cM. This finding helped to identify validated candidate genes involved in lipid biosynthesis and mapped near significant QTLs for various economic yield traits. This indicates the applicability of markers for MAS to improve the required trait selection for the oil palm breeding programs.

Moreover, Pootakham *et al.* (2015) developed SNP markers using the GBS method in the African oil palm with a total of 1085 SNPs to construct a linkage map. The map produced spanned 1429.6 cM and had an average of one marker every 1.26 cM. They also detected on LG 10, 14, and 15, three QTL genes affecting trunk height whereas a single QTL associated with fruit BW was identified on LG 3. They concluded that mapping the oil palm genome by the use of Genotyping by sequencing (GBS) approach helped to produce high-density maps and could enhance knowledge on genome structure which is valuable for mapping other economically important genes for MAS. Bai *et al.* (2018b) also used high-density GBS marker data to construct and detect QTL associated with leaf area using 145 oil palm breeding

populations derived from a cross between *Deli dura* and *Avros pisifera*. They constructed a genetic linkage map using a total of 2413 SNPs, producing a total of 16 LGs with a total length of 1161.89 cM, and an average marker spacing of 0.48 cM. Based on their results, two potential QTLs for leaf area were detected on Chr 3 and 9 and the gene ARC5, located in the QTL region on Chr 9, was the most likely candidate gene responsible for leaf growth in oil palm. They concluded that the use of a high-quality and SNP-based map supplies a base to fine map QTL for agronomic traits and MAS yield improvement in oil palm.

Furthermore, Gan *et al.* (2018) reported the first DArT-based genetic linkage maps using two closely related oil palm populations. For this purpose, they used a total of 1399 DArT and 1466 SNP markers. They produced a total of 16 major independent LGs with map lengths of 1873.7 and 1720.6 cM and with an average marker density of 1.34 and 1.17 cM, respectively. The integrated map was 1803.1 cM long with 2066 mapped markers and an average marker density of 0.87 cM. In conclusion, the use of the high-density marker DArTseq marker helped to generate high-density genetic maps in oil palm, and the integration of maps was also useful to study QTL analysis of important yield traits and other MAS studies. By the same token, Ong *et al.* (2019) also reported a linkage-based genome assembly in oil palm. To construct the map they used a total of 27,890 SNPs markers and generated a total of 16 LGs with a total map length of 1,151.7 cM and an average mapping interval of 0.04 cM. This mapping helped to study QTLs in sugar and lipid biosynthesis pathways. It also helped to improve knowledge of the current physical genome of commercial oil palm. Very recently SPET markers were used to construct a high-density genetic linkage map from a controlled cross of two oil palm genotypes (Herrero *et al.*, 2020). To construct the map, they used a total of 3,501 SPET markers with a total length of 1,370 cM and 1.74 markers per cM (0.57 cM/marker). This resulted in a total of 16 LGs with a total of 1,054 loci. From their work, they concluded that the application of these cost-efficient SPET markers is suitable for linkage map construction in oil palm and probably, also in other species.

CHAPTER II. MATERIAL AND METHODS

CHAPTER II. MATERIAL AND METHODS

II. 1. MATERIAL

II.1.1. Basic molecular data

The molecular data of the study were made up of DNA. They were granted through an official agreement between PalmElit (www.palmelit.com), CIRAD (www.cirad.fr), and EU-GENES (Intra-Africa Academic Mobility Scheme of the European Union). DNA was issued from plants located in North Sumatra, on the SOCFINDO estate (Indonesia), geographically between 2° 39' North - 99° 42' East at AekLoba Timur (ALT) and 2° 38' North - 99° 37' East at AekKwasan (AK) in North Sumatra at an altitude of 50 m above sea level with a distance of around 9 km between them (Fig. 12). The soil of the study area was characterized by deep well-drained soils developed over reworked Toba Tuffs (Cros *et al.*, 2017). In relation to this, some of the plant material was also located in Benin, on the INRAB research station of Pobè (Cros *et al.*, 2017).

The experiment used the standard trials for the evaluation of oil palm parental populations and laid out in Randomized Complete Block Designs (RCBD) with five or six blocks and/or in balanced lattices of rank four or five. Site ALT was established with 28 trials, AK was divided into AK1 with several trials, and AK2 composed of 19 trials (Fig. 13). All agronomic practices were applied based on the recommendation for oil palm crops (Cros *et al.*, 2017).

The plant material used in this experiment consisted of individuals from Group A and Group B are the two parental and heterotic groups involved in oil palm hybrid cultivar development (Ithnin & Din, 2020). Deli parental population belonging to Group A was derived from four individuals planted in 1848 in Indonesia (Hartley, 1988). This group also included individuals from the Angola population which resulted from material collected before the 1950s (Corley & Tinker, 2016). Group A produces a small number of large bunches while Group B produces a lot of small bunches. Group B is made up of several breeding populations mainly originating from Africa. La Mé population originated from Côte d'Ivoire (Corley & Tinker, 2016; Ithnin & Din, 2020).

It comprised 943 genotyped individuals with 423 Deli, 140 La Mé and 380 Deli × La Mé hybrid individuals (Table I). The Deli and La Mé populations used here were complex, involving several families with varying sizes and levels of relatedness. Thus, the Deli individuals belonged to 89 families of full-sibs with a mean size of 4.8 individuals (ranging from one to 60 individuals). The La Mé individuals belonged to 24 families of full-sibs with a mean size of 5.8 individuals and ranging from one to 31 individuals (Cros *et al.*, 2017).

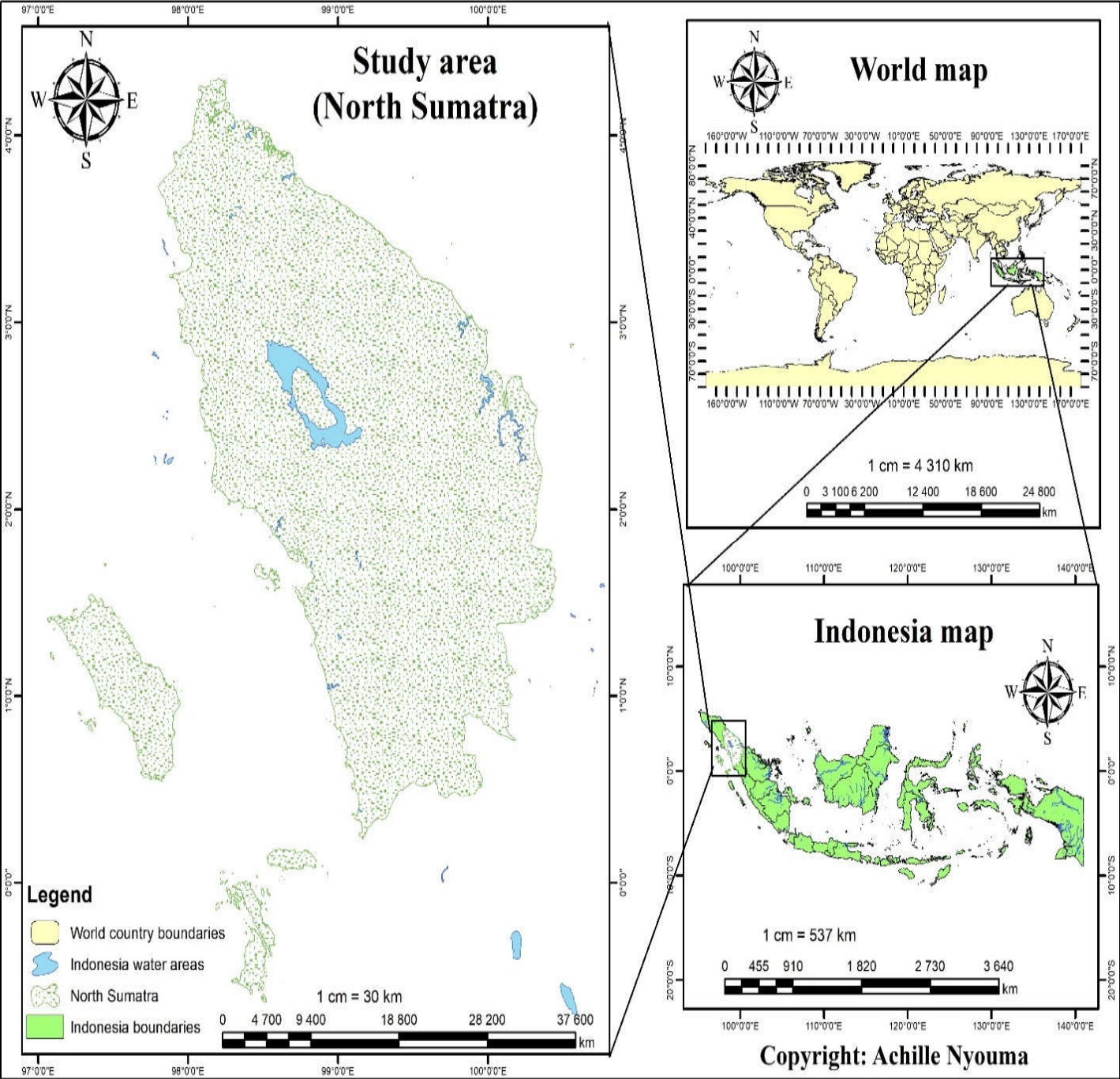


Fig. 12. Description of the location of plants used (Nyouma, 2021).

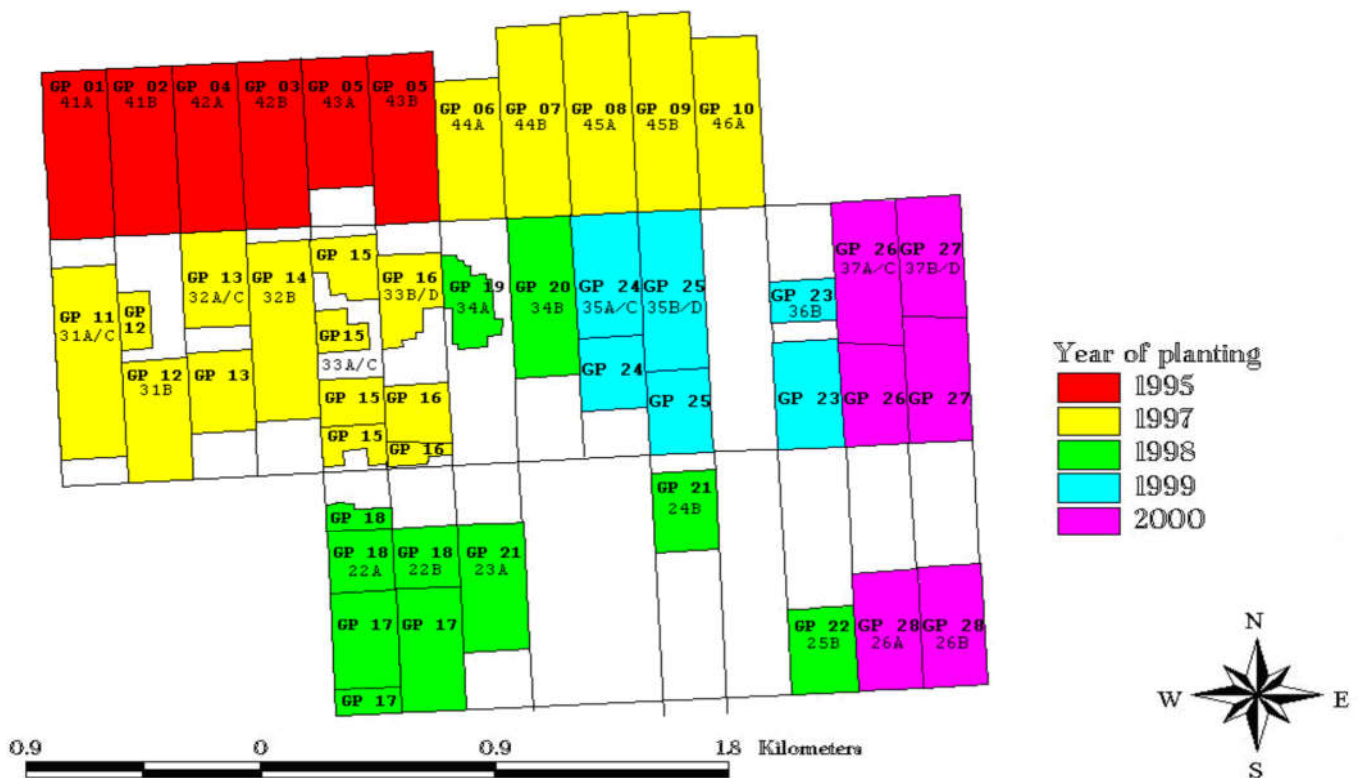


Fig. 13. Location plan of the 28 trials (GP) of AekLobaTimuer (Cros *et al.*, 2014).

Detailed pedigree information of these two populations is known over several generations (Cros *et al.*, 2017). The Deli \times La Mé hybrid individuals were obtained by crossing 67 and 63 of these Deli and La Mé individuals, respectively, according to an incomplete factorial design. The hybrid individuals belonged to 101 crosses comprising on average 3.8 individuals (ranging from one to 10). For the construction of the genetic map, all the genotyped Deli, La Mé and Deli \times La Mé individuals were used, as well as the non-genotyped individuals comprised in their pedigree, for a total of 1,788 individuals. For the other parts of the study, only genotyped individuals of the Deli and La Mé breeding populations were used (Table I).

Table I. Oil palm plant material used.

Breeding populations	Total number of individuals	Number of genotyped individuals	Number of full-sibs families	Mean number of Individuals per family
Deli	423	423	89	4.8 (1-60)
La Mé	140	140	24	5.8 (1-31)
Deli \times La Mé	388	380	110	3.8 (1-10)
Total	951	943	-	-

II.1.2. Other material

A computer with the most important features x 64-based processor, 64-bit operating system, Intel(R) Core(TM) i7-8550U CPU running at 1.80GHz or 1.99GHz, and 16.0 installed memory (RAM) was used. MobaXterm (<https://mobaxterm.mobatek.net/>), FileZilla (<https://filezilla-project.org/>), WinSCP (<https://winscp.net/eng/index.php>), and 7-Zip (<https://www.7-zip.org/>) were the key software used for each analysis in this experiment. The server for large molecular data analysis was accessed through the IFB Core Cluster server (<https://my.cluster.france-bioinformatique.fr>).

II.2. METHODS

For this specific study, the DNA extraction had been carried out at the laboratory of ADNid (www.adnid.fr) France using lyophilized tissue from the youngest opened leaf of each individual, using a modified mixed alkyltrimethylammonium bromide (MATAB) protocol and more detail of the procedure was found (Cros *et al.*, 2017). For the subsequent studies, the final generated molecular data was common to all, and more detail about this activity was elaborated on in the following paragraph.

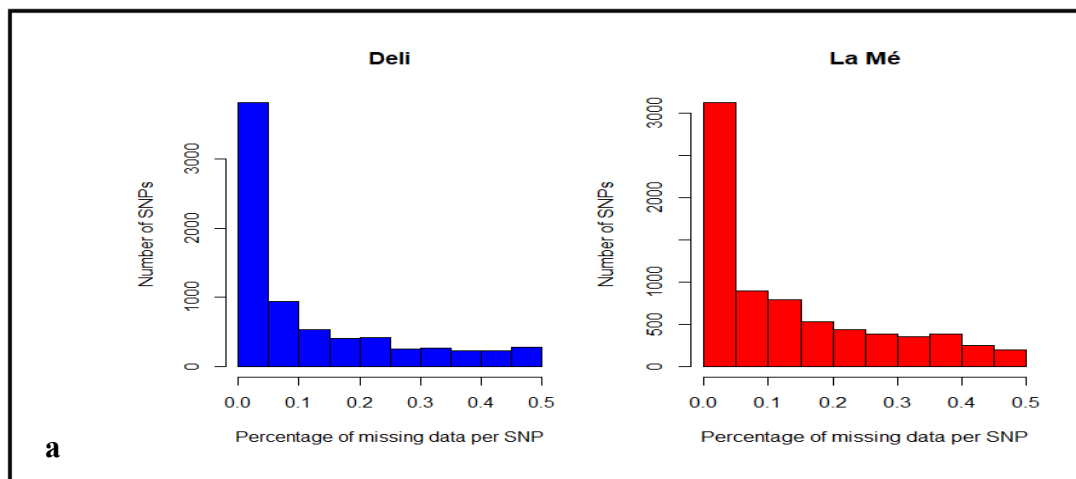
II.2.1. Generation of molecular data

Molecular data were obtained by using genotyping by sequencing (GBS) (Appendix 4) (He *et al.*, 2014). GBS and SNP calling were performed based on the procedure described (Cros *et al.*, 2017). A total of 96 plant sample kits of DNA were used and it undergoes digestion/ligation reactions by using two different adaptors namely *PstI* and *HhaI*. The *PstI* adaptors included a sequencing primer so that the sequences were always read from the *PstI* restriction sites, the sequencing primer sequence, and the “staggered”, varying length barcode region (Elshire *et al.*, 2011). In 30 rounds of PCR, only *PstI*-*HhaI* mixed fragments were successfully amplified under the following 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. The amplification products from each sample in the 96-well microtiter plate were then bulked up and used in c-Bot bridge PCR (Illumina) before being sequenced on Illumina HiSeq2500. There were 77 cycles of single-read sequencing (Cros *et al.*, 2017).

The sequence data were processed using Tassel GBS version 5.2.44 (Glaubitz *et al.*, 2014) and VCFtoolsver 0.1.14 (Danecek *et al.*, 2011). The reference genome of Singh *et al.*

(2013) was used for alignment with Bowtie2 software (Langmead & Salzberg, 2012). Biallelic SNPs were the only variants kept. SNP data points with a depth below 10 were set to missing and only SNPs with less than 50% missing data in the two breeding populations were kept. SNPs with a sum of depth per datapoint above 550,000 and SNPs with 100% heterozygote genotypes were discarded because it was considered that this might be a sign of genome duplication. Individuals with more than 50% missing data were removed. Finally, 7,324 SNP markers were obtained, common to both breeding populations, including 5,598 SNPs located on the assembled parts of the genome due to the need for known positions in order to impute sporadic missing data, i.e. the 16 chromosomes (Singh *et al.*, 2013). Two copies of the molecular dataset were created, one for Group A (i.e, Deli) and the other for Group B (i.e, La Mé) by using the pedigree file information Cros *et al.* (2017) and VCFtoolsver 0.1.14 software (Danecek *et al.*, 2011).

The average percentage of missing data per SNP was 11% in Deli, and 13% in La Mé (Fig. 14a). Also, the average percentage of missing data per individual was 11% in Deli, and 13% in La Mé (Fig. 14b). Again, the average percentage of heterozygosity per SNP was 7 % in Deli and 10% in La Mé (Fig. 14c).



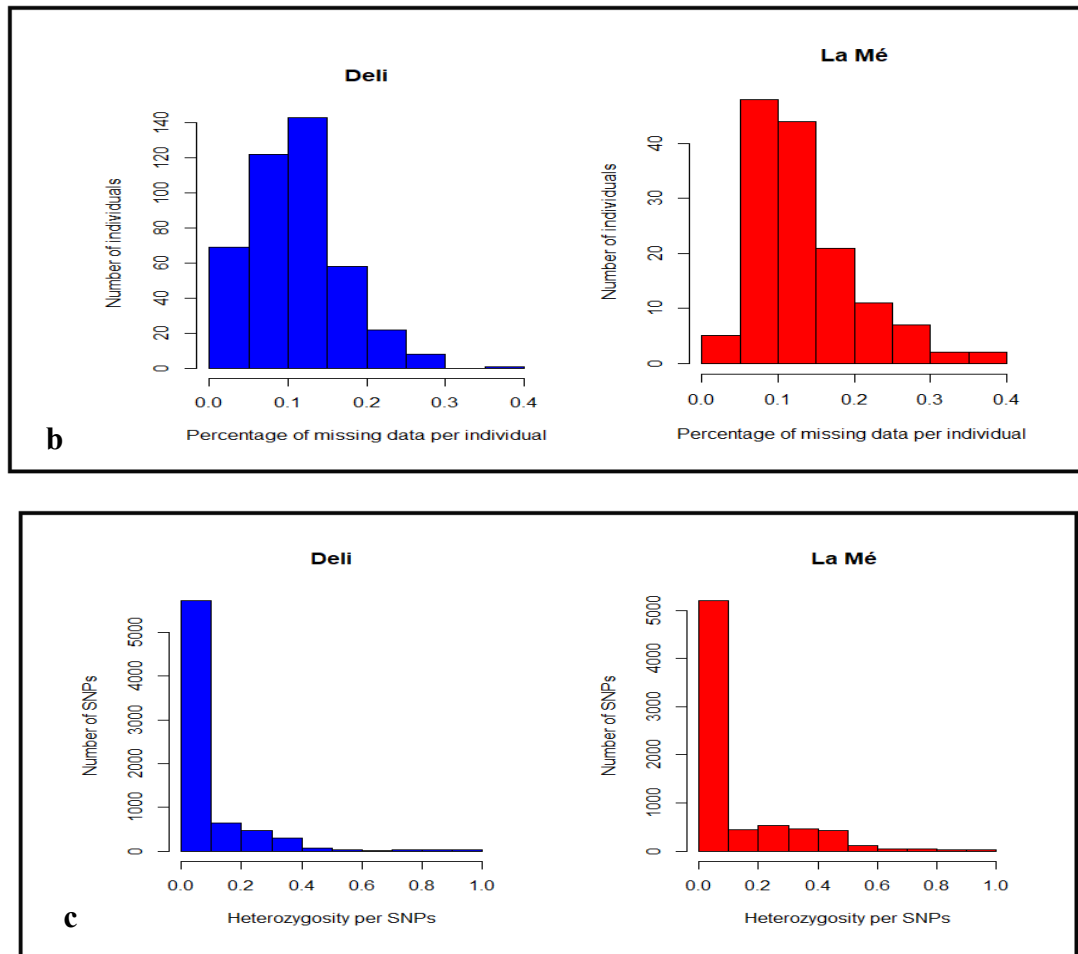


Fig. 14. Distribution of percentage of missing data for oil palm breeding populations per SNP (a), per individual (b), and percentage of heterozygosity (c).

II.2.2. Genome mapping

To calculate the total number of markers required for better genomic selection and genomewide association studies it is very critical to know about the actual position of markers in the genetic position (cM). For this purpose, the genetic maps were constructed to know the actual position of markers that are found in the genetic positions in both Deli and La Mé of oil palm breeding populations, and the finding of this result was also used to calculate the within-population LD, the persistence of LD and haplotype sharing between the two breeding populations in the genetic distance (cM) and detail procedure for the construction of genetic linkage map was outlined in the following paragraph below.

II.2.2.1. Construction of the genetic maps

Genetic maps were made using LepMAP3 software (Rastas, 2017). First, module “ParentCall2” was used to call missing or erroneous oil palm genotypes, with parameters “removeNonInformative=1” and “halfSibs=1”. Secondly, the “Filtering2” module handled the

filtering of the data for marker quality checking. In this module, markers that were monomorphic or homozygous in both parents were removed and segregated in a non-Mendelian fashion using “removeNonInformative=1” and “dataTolerance=0.001”. Thirdly, the “SeparateChromosomes2” module assigned markers into LGs by computing all pair-wise LOD scores between markers and joined markers with LOD scores higher than the user-given parameter “LodLimit”, which was set to 8. Fourthly, the “JoinSingles2All” module assigned singular markers to the existing LGs by computing LOD scores between every single marker and markers from the existing LGs, using “numMergeIterations=10” and “numThreads=12”. Finally, “OrderMarkers2” ordered the markers within each LG by maximizing the likelihood of the data given the order and the Kosambi mapping function for conversion of recombination frequencies into map distances in centiMorgan, cM (Rastas, 2017). To join the maps of both male and female parents, the sex average argument was set to 1. The individuals that were associated with outlier values in terms of the number of crossing-overs were removed. The markers which created large gaps at the top or bottom part of LGs were also canceled. The LGs with a low number of SNPs were discarded to keep a genetic map with the number of LGs corresponding to the number of chromosomes of oil palm, i.e. 16.

II.2.2.2. Comparison of genetic and physical maps

The genetic map and physical maps, showing the positions of the reference genome of Singh *et al.* (2013), were visualized using the R package LinkageMapView (Ouellette *et al.*, 2018). MareyMap of Siberchicot *et al.* (2017) to plot the genetic positions of the molecular markers against their physical position were used.

II.2.2.3. Comparative genomics

The composition and the position of SNPs originating from old reference genomes i.e., the Eg5.1 genome of Singh *et al.* (2013) with the newly published references genome i.e., PMv6 Ong *et al.* (2020) were compared. To do so, MareyMap, as described by Siberchicot *et al.* (2017) to plot the Eg5.1 genome positions against their PMv6 genome position, were used and the percentage of the repositioned SNPs with the old SNPs that were carried by the same chromosome in the two genomes were also computed.

II.2.3. Evaluation of genetic diversity of Deli and La Mé populations

II.2.3.1. Allele and genotype frequencies

The distribution of MAF, percentage of heterozygosity per individual, correlation of heterozygosity per SNPs, and the frequency of alternate alleles per SNP both in Deli and La Mé of oil palm breeding populations were analyzed using the 7,324 SNPs generated and available in R software (R Development Core Team, 2022).

II.2.3.2. Fixation index (F_{st})

The pairwise Fixation index (F_{st}) between Deli and La Mé of oil palm breeding populations was estimated according to Wright (1931), using the 7,324 SNPs available and subsets of 100 random individuals per population to avoid a biased in computing the F_{st} values between an unequal number of genotyped individuals per population (Gondro *et al.*, 2013). The Fixation index (F_{st}) value was obtained using the SNPRelate R package (Zheng *et al.*, 2012).

II.2.4. Estimation of within-population linkage disequilibrium

Analyses of LD were performed in each breeding population (i.e., Deli and La Mé) using the PLINK software (Purcell *et al.*, 2007). It computed pairwise estimates of LD by the classical measure of the squared correlation of allele frequencies at diallelic loci (r^2) and r . Before the computation of the r^2 , the missing data points in the Deli and La Mé individuals were imputed using Beagle 5.2 Browning *et al.* (2018), independently for each breeding population (Appendix 5).

For the SNPs located on the assembled parts of the genome, the r^2 values between pairs of SNPs were plotted against physical distances (Mbp). For the SNPs located on the genetic map, the r^2 values were plotted against genetic distances (cM). The LD decay was plotted up to a 0.8 Mbp distance for physical positions and 3 cM for genetic positions. The relation between the r^2 values and distances was modeled by fitting local polynomials with the functions ‘locpoly’ and ‘dpill’ of the R package KernSmooth Version 2.23 Wand (1995), as done for example in (Yamamoto *et al.*, 2016).

The persistence of LD between populations was measured by the correlation of the r measure of LD between populations given by PLINK (r_{LD}). The r_{LD} was computed between the two populations on the SNPs comprised in windows defined along with the genetic and physical maps, over a distance up to 90 cM and 50 Mbp, respectively. The r_{LD} values can vary from -1

to 1, with a value close to 1 indicating a similar LD pattern in the two populations for the SNPs located in the genomic window considered.

II.2.5. Assessment of haplotype sharing of Deli and La Mé population

The percentage of shared haplotypes between the Deli and La Mé oil palm breeding populations was analyzed according to the length of the genomic window represented in both genetic and physical distances using the SNPs located on the assembled part of the genome. The SNPs data were phased using Beagle 5.1 (Browning *et al.*, 2018). The phasing of the SNPs data was undertaken by considering the two oil palm breeding populations independently.

Sliding windows were defined along the chromosomes and LGs, with an overlap of 50%. Fifteen window sizes were used for physical distances, from 10 Mbp to 100 bp, and seven window sizes were used for genetic distances, from 10 cM to 0.01 cM. The window sizes were considered by decreasing order and, for each window of a given window size, the list of haplotypes existing in each population was made after discarding the haplotypes with the actual length shorter than the next window size.

Upon the analysis to avoid redundancy that could result from the overlap between windows, only a single copy of the duplicated haplotypes (i.e. haplotypes identical in sequence and starting at the same position) was kept. Finally, the length of the haplotypes, the percentage of haplotypes common to the two populations, and, for the common haplotypes, their frequency in each population were computed. This analysis was done using custom R software (R Development Core Team, 2022).

II.2.6. Determination of the effective population size of Deli and La Mé

The effective population size (N_e) of the both Deli and La Mé oil palm breeding populations was calculated using multi-locus samples obtained from a total of 7,324 SNP markers. We analyzed the effective population size of Deli and La Mé breeding populations independently and upon analysis, the issue of missing data from the molecular data, screening out of the rare alleles from molecular data, and problems related to sampling individuals from the population were controlled by NeEstimator software and an equal amount of individuals were taken from each breeding population (Do *et al.*, 2014).

The effective population size was estimated with the LD method based on the linkage of Waples & Do (2008) implemented in the NeEstimator 2.1 software (Do *et al.*, 2014). The

computation was made separately for the Deli and La Mé oil palm breeding populations using the SNPs located on the genetic map and the assumption of random mating. The confidence interval of effective population size values was obtained by the Jackknife method on samples (Waples & Do, 2008).

CHAPTER III. RESULTS AND DISCUSSION

CHAPTER III. RESULTS AND DISCUSSION

III.1. RESULTS

III.1.1. Genetic diversity of Deli and La Mé

III.1.1.1. Distribution of minor allele and genotype frequencies across the population

The distribution of MAF in the Deli and La Mé oil palm breeding populations showed a reduction in the number of SNPs with the increase in MAF (Fig. 15). The MAF ranged from 0.0 to 0.5 for both Deli and La Mé oil palm breeding populations. Thus, the average MAF was 0.09 for Deli and 0.14 for La Mé. In both breeding populations, most SNPs had low MAF <0.05 values. Thus, the percentage of SNPs with MAF <0.05 was 60.5% in Deli, and 49.7% in La Mé (i.e., around 10.8% more SNPs with low MAF in Deli than La Mé). This demonstrated that Deli parents had less genetic diversity than La Mé parents.

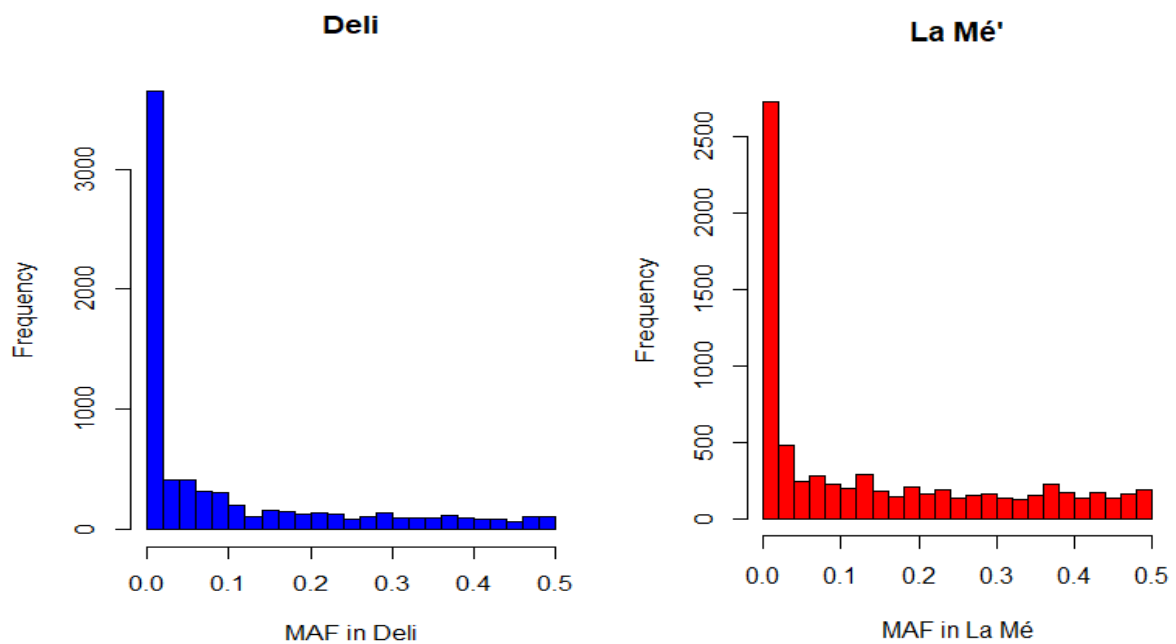


Fig. 15. Distribution of minor allele frequency (MAF) in Deli and La Mé oil palm breeding populations.

III.1.1.2. Heterozygosity

The percentage of heterozygosity per individual ranged from 0.00 to 0.20 for both breeding populations. When the percentage of heterozygosity per individual increases the number of heterozygote individuals decreases for both Deli and La Mé populations. Thus, the percentage of heterozygosity per individual ranged from 1.9% (Deli) to 20.9% (La Mé). Deli was the population with the lowest percentage of heterozygote SNPs (mean 7%), while La Mé has a higher level of heterozygosity (10%) (Fig. 16).

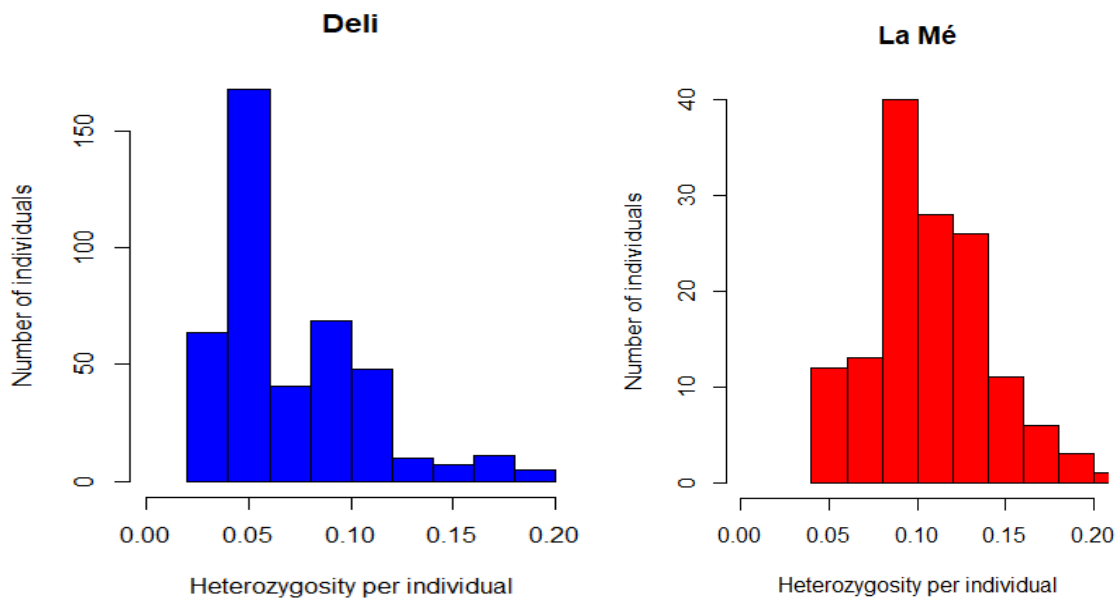


Fig. 16. Distribution of the percentage of heterozygosity per individual for Deli and La Mé oil palm breeding populations.

Accordingly, some Deli individuals were up to 0.16% homozygotes, while for La Mé 0.20% were homozygotes. More than 150 individuals are less heterozygosity percentage (7%) in Deli than in the La Mé population (> 40 individuals) with a heterozygosity percentage value of 10%. In the same view, the distribution of the percentage of heterozygosity per individual also showed that the majority of individuals are less heterozygosity in Deli than in La Mé populations. This indicated that the Deli populations are more homozygote than the La Mé populations.

The correlation of heterozygosity per SNPs among Deli and La Mé oil palm breeding populations (Fig. 17) showed that the majority of SNPs are fixed or almost fixed (i.e.

concentrated alongside the x and y axes) either in Deli or La Mé population while, in the other population, they had a much larger level of heterozygosity.

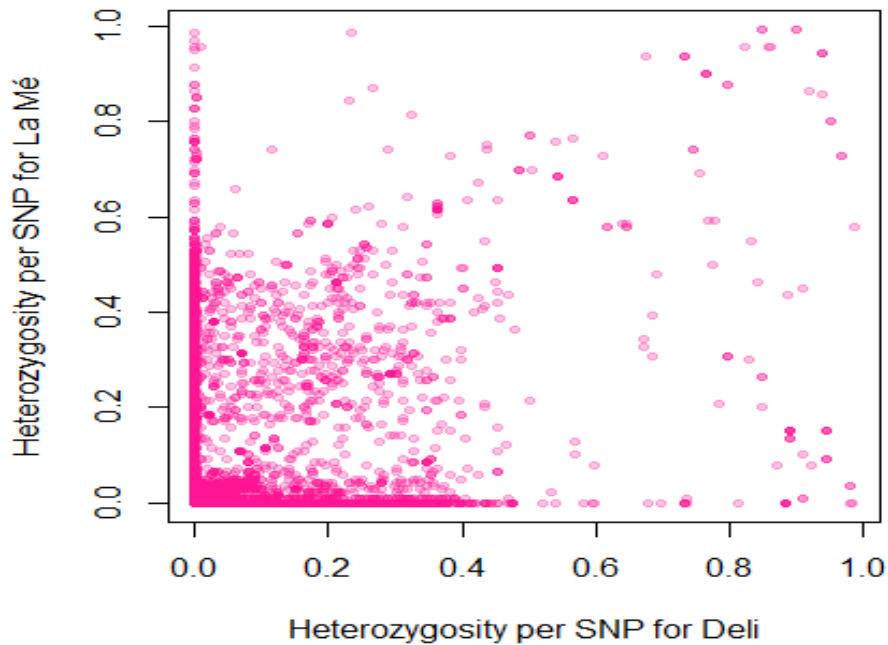


Fig. 17. Correlation of heterozygosity per SNPs among Deli and La Mé in oil palm breeding populations. Each dot represents an SNP.

There was a higher heterozygosity percentage between Deli and La Mé oil palm breeding populations. Between the two populations, the central part of the plot had the smallest number of SNPs, indicating a strong genetic divergence between them. This indicated that Deli, the southeast Asia origin is closer to oil palm originating from central African origin oil palm populations than the WA oil palm origin i.e., La Mé.

The correlation in the frequency of alternate alleles per SNPs among the populations showed that SNPs largely concentrated alongside the x and y axes, demonstrating that most SNPs have distinct segregation patterns among populations (Fig. 18).

The SNPs were being fixed or almost fixed in one individual while segregating in the other individuals. A large proportion of SNPs thus appeared fixed with one allele in one population and with the other allele in the other population. In both breeding populations, alleles are fixed to either the x and y axes and less distributed in the middle which indicates that the two breeding populations are genetically divergent.

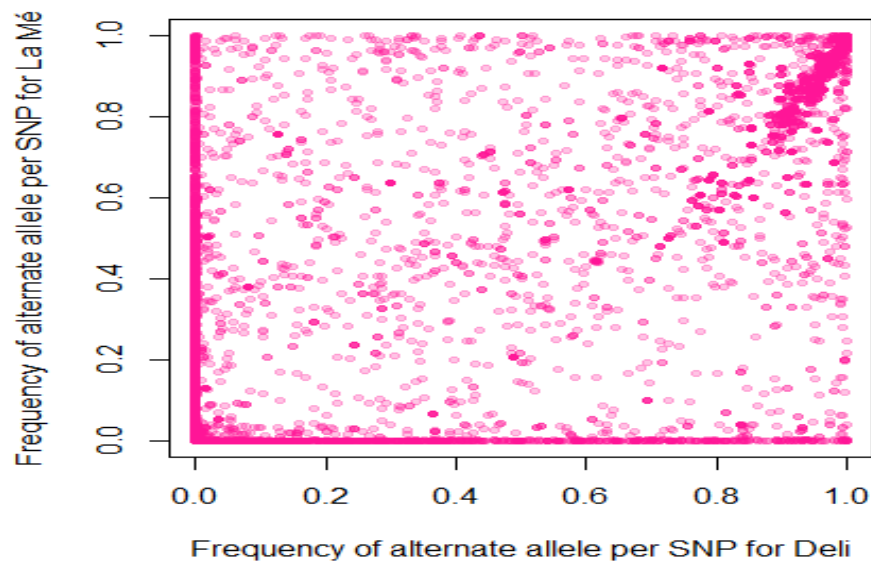


Fig. 18. Correlation of frequency of alternate allele per SNPs among Deli and La Mé in oil palm breeding populations. Each dot represents an SNP.

III.1.1.3. Fixation index (F_{st})

The result of the fixation index showed the degree of differentiation among Deli and La Mé oil palm breeding populations which indicates that there is a high genetic divergence among these two breeding populations. There was the highest degree of differentiation between them with an F_{st} value of 0.53. Data in Fig. 19 showed the Fixation index between populations at the chromosome level. Several regions of the genome had high F_{st} values (> 0.6), in particular on chromosomes EG51_1, EG51_8, and EG51_13. Depending on the region of the genome considered, there were large variations in the magnitude of the differences in fixation index among the two pairs of populations, and their rank often differed. There was a maximum average peak value between the two populations with a 16.61 mean peak value. Thus, a peak observed for Deli and La Mé on chromosomes EG51_1, EG51_2, and EG51_10 did not collocate with peaks in other pairs of populations.

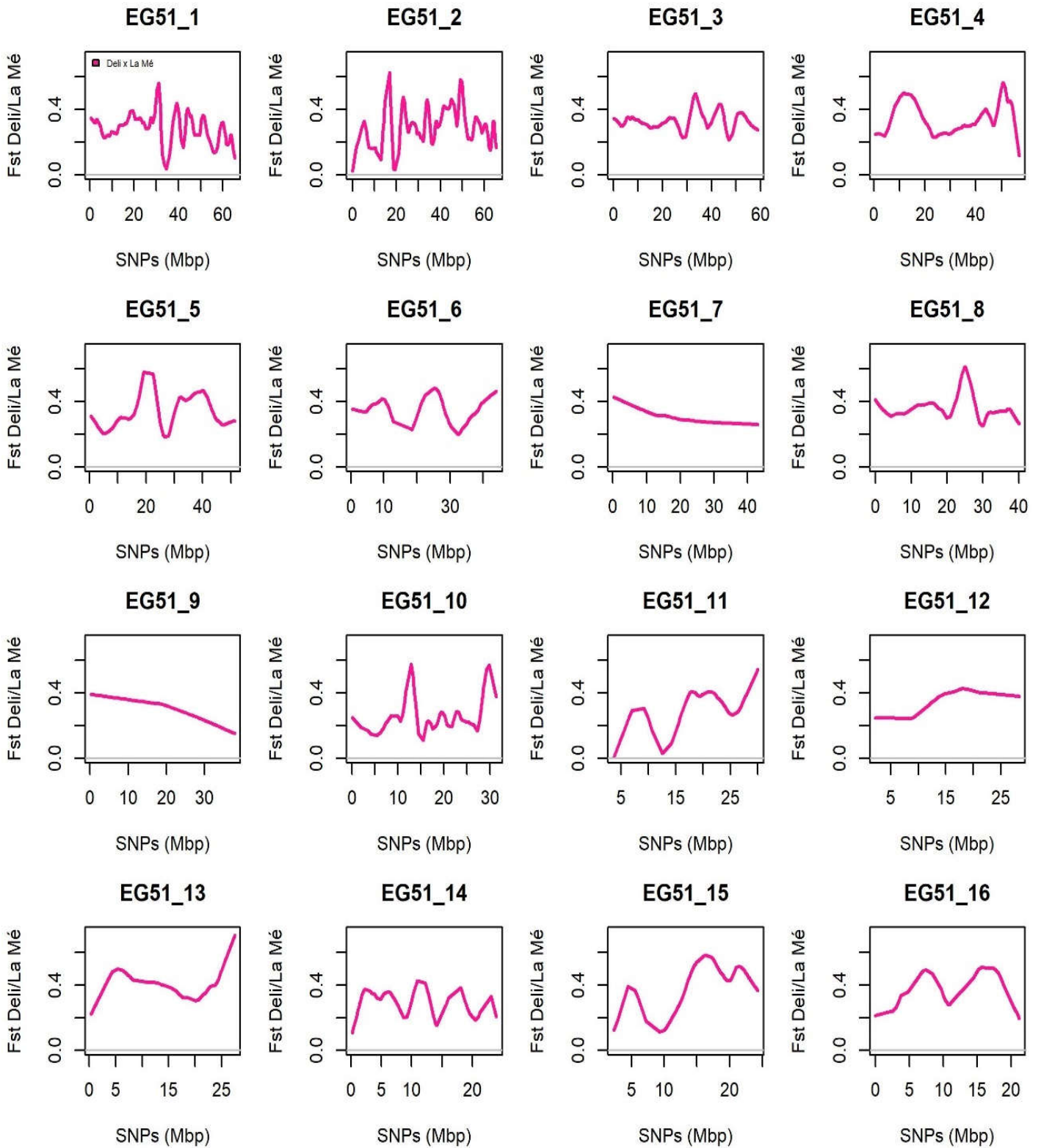


Fig. 19. Fixation index value between Deli and La Mé oil palm breeding population along with physical distances (Mbp). * EG: *Elaeis guineensis* Mbp: Megabase pair

III.1.2. Within-population linkage disequilibrium of breeding populations

The analysis revealed a rapid decrease in the average genome-wide pattern of LD in both Deli and La Mé breeding populations with increasing genetic (cM) and physical (Mbp) distances (Fig. 20 and Fig. 21).

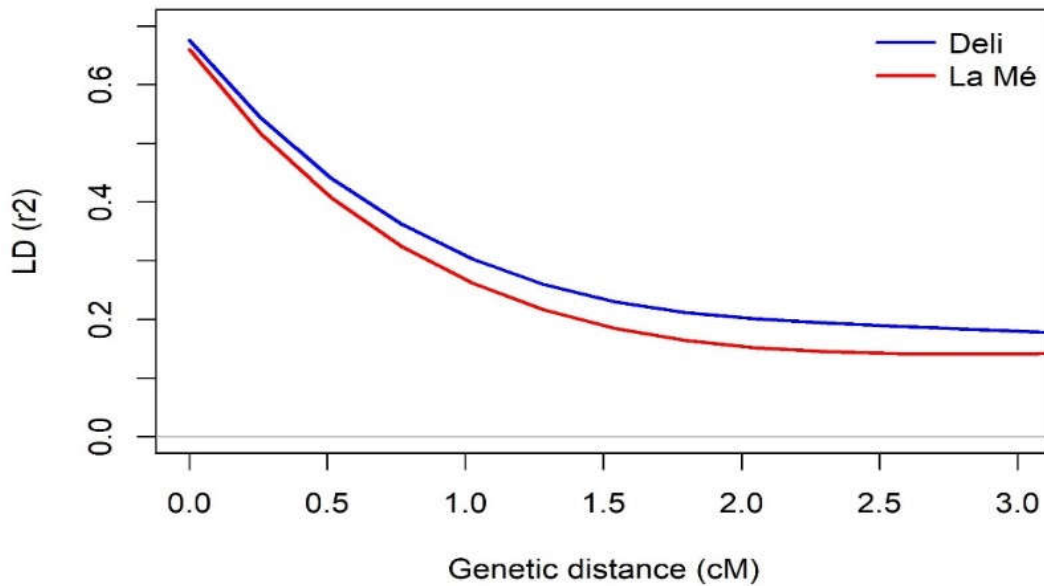


Fig. 20. Average genome-wide pattern of linkage disequilibrium decay between pairs of SNPs (r^2) according to the genetic distance (cM) between SNPs. *cM: centiMorgan

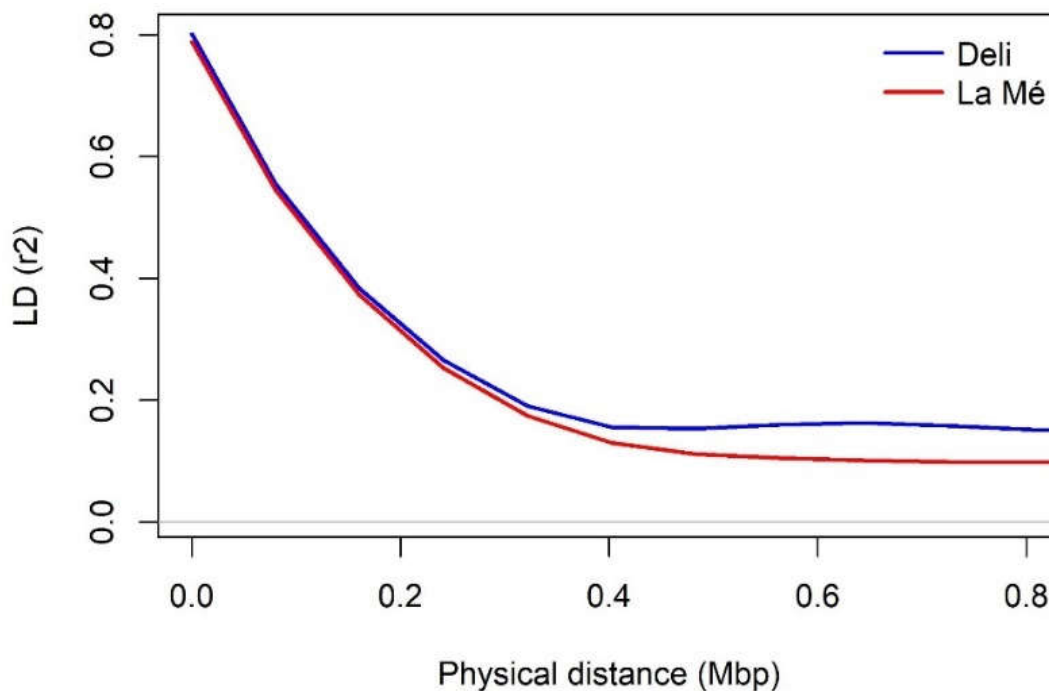


Fig. 21. Average genome-wide pattern of linkage disequilibrium decay between pairs of SNPs (r^2) according to the physical distance (Mbp) between SNPs. *Mbp: Megabase pair

The decay of LD between pairs of SNPs according to the genetic distances is shown in (Fig. 20). The LD reached high values (> 0.6) for short distances between SNPs. It was higher in the Deli population than in the La Mé population for all distances. For example, considering the r^2 value of 0.3, the corresponding distance between SNPs was 1.05 cM in Deli and 0.9 cM in La Mé. The difference between the two populations was small for short distances and increased with the distance between markers.

Similar trends were observed when plotting LD against physical distances (Mbp) (Fig. 21), although the r^2 values reached higher levels (i.e. around 0.80), as a consequence of the higher number of markers on the physical map than on the genetic map. The distance corresponding to $r^2 = 0.3$ was 0.22 Mbp in Deli and 0.21 Mbp in La Mé.

III.1.2.1. Persistence phase between Deli and La Mé populations

A strong and high persistence of phase correlation of r_{LD} values between Deli and La Mé populations was observed for close markers. Phase correlations decreased rapidly with increasing distances between SNP, as was similarly observed for average r^2 in both Deli and La Mé breeding populations. Phase correlations (r_{LD}) above 0.6 for SNPs separated by a distance < 0.5 cM on the genetic map or < 1 kbp on the physical map (Fig. 22). The r_{LD} value decreased sharply with the distance between SNPs increased, and was thus divided by two before 2 cM and 5 Mbp, and became negligible at distances above 50 cM or 50 Mbp.

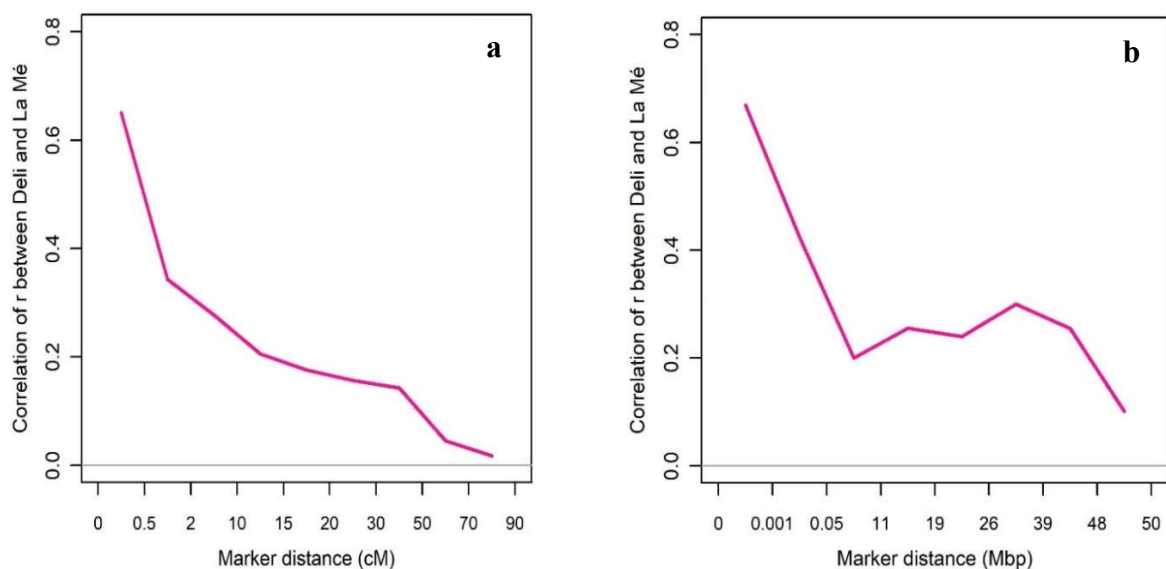


Fig. 22. Correlation of the r measure of LD between populations as a function of genomic distance in cM (a) and Mbp (b). The dotted grey line indicates $r = 0.1$.

III.1.2.2. High-density genetic map

Of the 7,324 SNP markers in the final VCF file for genome properties, 4,252 SNP non-redundant markers were located on a genetic linkage map that was spread over 16 LGs (Fig. 23).

The LGs with a low number of SNPs were discarded to keep a genetic map with the number of LGs corresponding to the number of chromosomes of oil palm.

The genetic map comprised 4,252 SNPs, spread over 2,782 unique positions (Table II, Fig. 23), and spanned 1,778.52 cM. Even coverage of the genome was achieved, with an average mapping interval between adjacent SNPs of 0.67 cM.

The number of unique SNP positions mapped to each linkage group ranged from 87 (LG14) to 358 (LG1), with a mean of 174.93 SNPs per linkage group. The biggest gap size between SNPs ranged from 3.31 cM (LG11) to 6.66 cM (LG14). The size of the LGs ranged from 215.72 cM to 64.75 cM (Table II).

The longest linkage group was observed for LG1 (215.72 cM) and the shortest linkage group was obtained at LG 16 (64.75 cM) also reflected a similar ranking in terms of their marker sizes (Table II). The number of SNPs in each linkage group is also shown on the genetic map for each linkage group (Table II, Fig. 23).

The average inter-marker distance for the linkage map was 0.67 cM with distances ranging from 0.50 cM (LG3) to 0.88 cM (LG9). From all, around 56% of the SNPs marker interval had less than the average marker distance. There were also small to large gaps between LGs, accordingly, the smallest gap was observed at LG 11 with a 3.31 cM gap and the longest gap was observed at LG 14 with a 6.66 cM gap.

The recombination rate was 2.85 cM/Mbp on average, ranging from 1.78 cM/Mbp (LG15) to 3.87 cM/Mbp (LG13) (Table II).

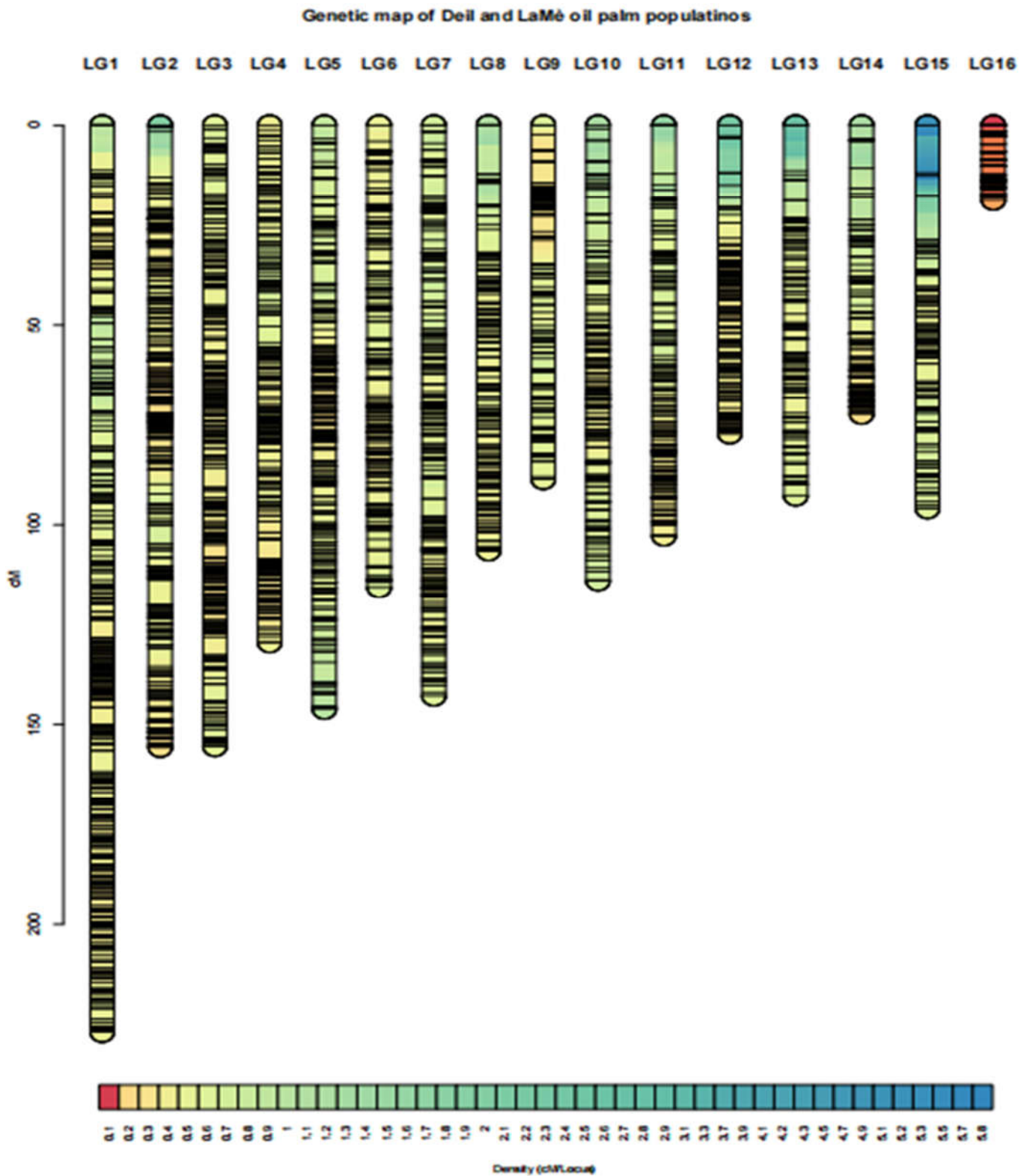


Fig. 23. Genetic map with 4,252 SNP markers. The y-axis indicates the distances (cM), and the colors indicate the density of markers according to the bottom scale (cM/locus).

*cM:centiMorgan; LG: Linkage group

Table II. Summary of the genetic map.

Linkage group	Number of markers	Length in cM	Average gap size (cM)	Biggest gap size (cM)	Number of unique positions	Corresponding chromosome (Singh <i>et al.</i> , 2013)	Number of common markers	Spearman correlation (absolute value)	Recombination rate (cM/Mb)
LG1	554	215.72	0.60	5.20	358	EG51_2	271	0.86	2.19
LG2	436	142.59	0.51	6.42	279	EG51_1	311	0.83	3.41
LG3	432	155.39	0.50	4.75	309	EG51_3	318	0.80	2.67
LG4	326	129.51	0.60	4.91	218	EG51_7	257	0.95	2.53
LG5	312	142.82	0.64	5.09	223	EG51_4	222	0.72	3.34
LG6	278	111.51	0.68	4.35	164	EG51_6	154	0.94	2.56
LG7	277	142.75	0.69	5.04	207	EG51_5	219	0.94	2.55
LG8	220	94.21	0.66	5.70	144	EG51_10	162	0.79	2.36
LG9	225	88.64	0.88	6.04	102	EG51_16	79	0.54	3.70
LG10	216	113.85	0.76	4.92	150	EG51_8	154	0.91	3.63
LG11	204	90.64	0.63	3.31	144	EG51_12	133	0.71	3.02
LG12	185	65.27	0.54	3.80	123	EG51_11	132	0.97	2.30
LG13	163	81.31	0.84	4.95	98	EG51_9	90	0.86	3.87
LG14	158	72.31	0.84	6.66	87	EG51_14	122	0.90	3.13
LG15	136	67.25	0.68	4.40	100	EG51_13	66	0.93	1.78
LG16	130	64.75	0.70	4.54	93	EG51_15	92	0.96	2.50
Sum	4,252	1,778.52			2,799		2,782		
Mean	265.75	111.15	0.67	5.00	174.93		173.875	0.85	2.85

*cM:centiMorgan; LG: Linkage group; Mb: Mega base

III.1.2.3. Comparison of genetic and physical maps

The physical and genetic orders were in general in agreement, with a Spearman rank correlation above 0.7 for 15 LGs out of 16. However, upturns of large chromosome segments between the genetic map and the reference genome existed in a few cases, for instance, in chromosomes 16, and, to a lesser extent, 1, 12, and 14 (Fig. 24, Fig. 25 and Table III).

On the other hand, chromosomes 1, 2, 11, 12, 13, and 16 had a slight to larger horizontal gap compared to the other chromosomes which indicates a low SNP coverage in the particular genomic regions. Generally, punctual disagreements between physical and genetic distances concerning a few SNPs appearing as outliers, i.e. far apart from the regression line, were also observed in all chromosomes except 7 and 13.

Data depicted in Table III and Fig. 25 showed a summary of the physical map that indicated the SNPs located on the assembled part of the genome. We obtained a total of 2,782 markers in the physical position with an average of 174 SNPs markers.

A larger number of SNPs was observed at chromosome EG51_3 (318) and the smallest was observed at chromosome EG51_15 (99). The physical map encompassed a mean of distance 40.10 Mbp, with chromosome distances ranging from 65.07 Mbp (EG51_1) to 21 Mbp (EG51_9).

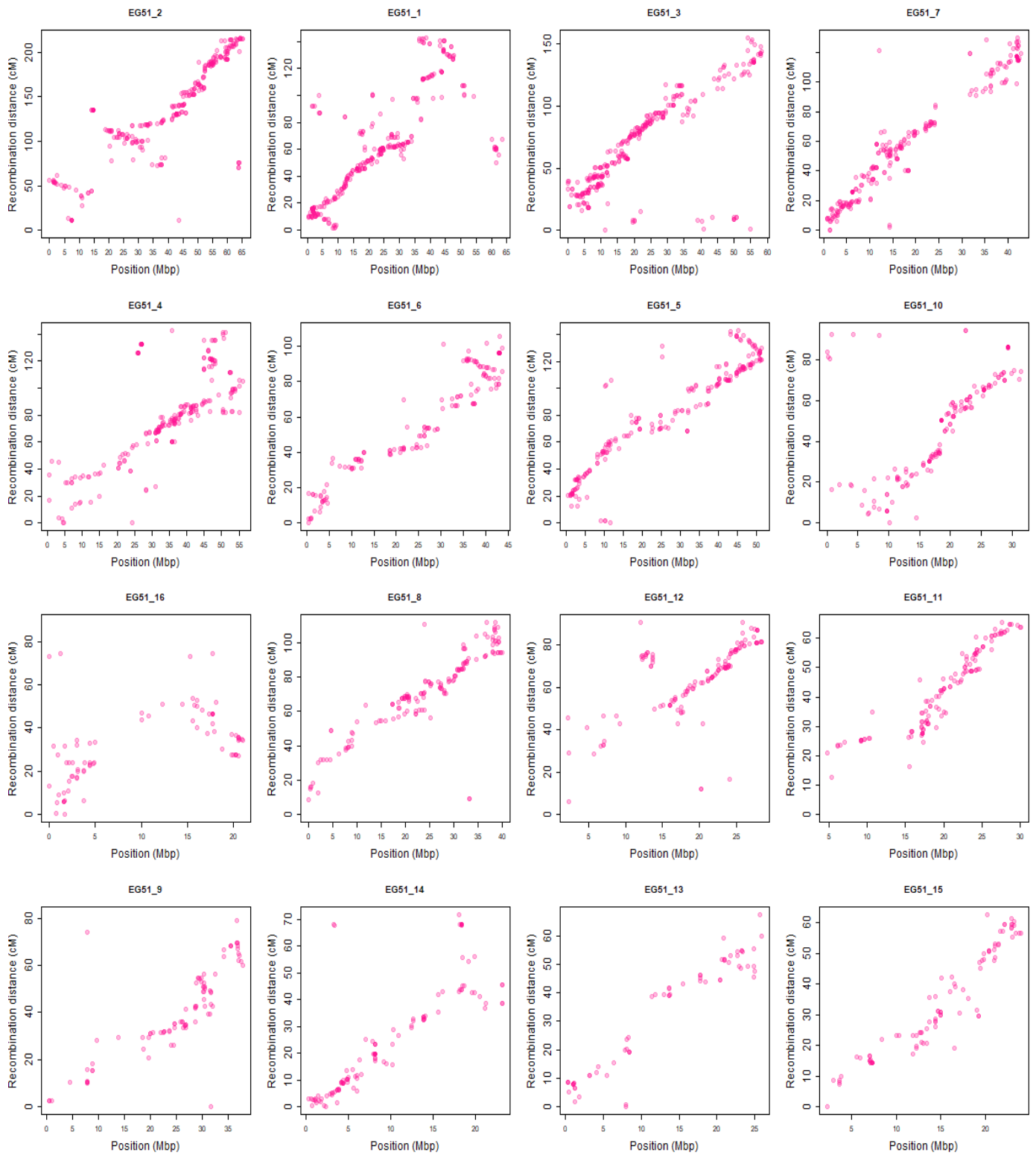


Fig. 24. Visualization of marker genetic positions (cM) versus physical positions (Mbp) for each chromosome.

*EG: *Elaeis guineensis*; Mbp: Megabase pair and cM: centiMorgan

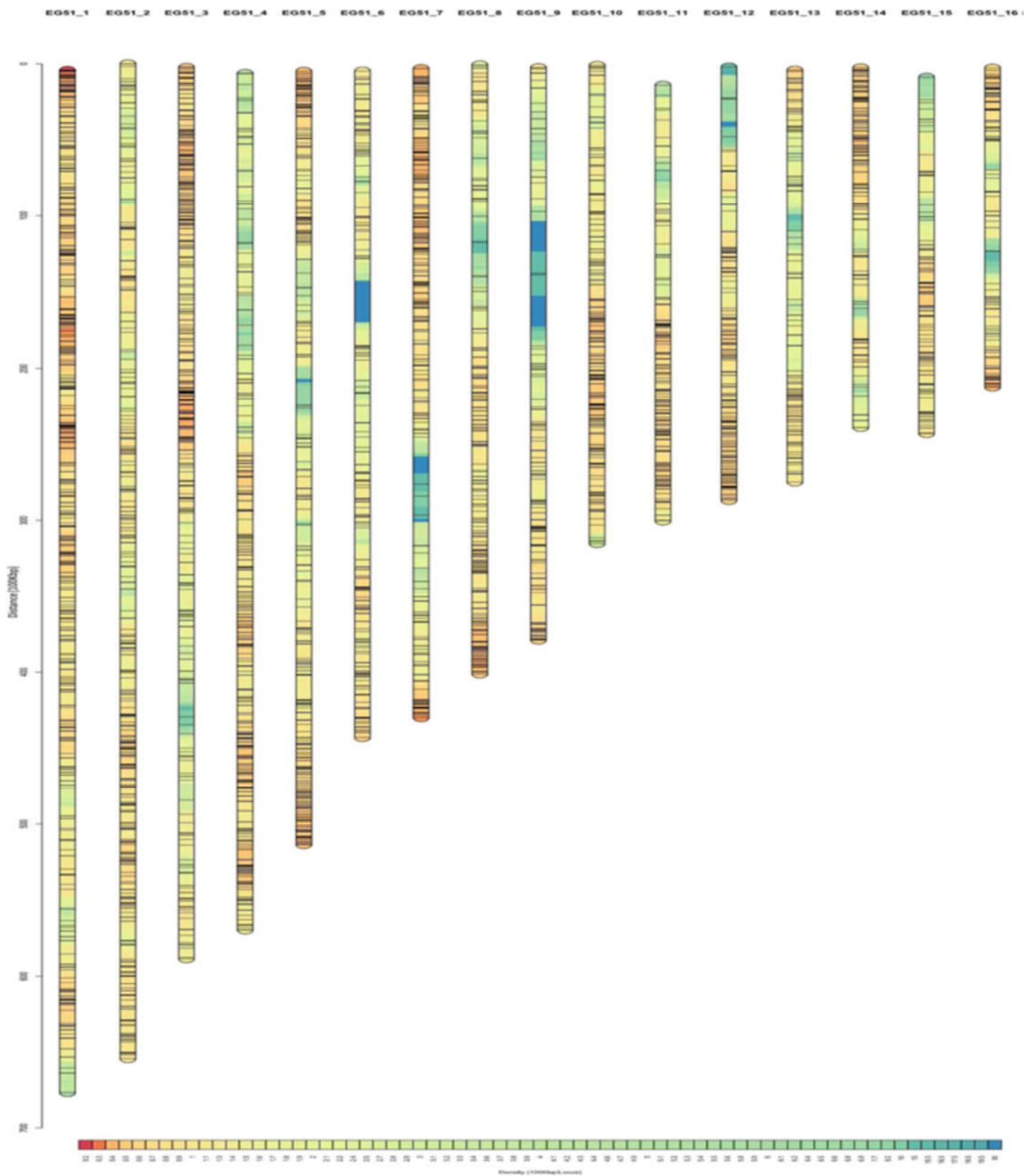


Fig. 25. Physical map of Deli and La Mé oil palm populations with 5,598 SNP markers. The colors indicate the density of markers according to the bottom scale (100 kbp/locus).

*EG: *Elaeis guineensis*; kbp: Kilo base pair

Table III. Summary of the physical map (SNPs located on the assembled part of the genome).

Chromosome Name	Number of Markers	Length (bp)	Average Distance of Markers (bp)	Maximum distance of Markers (bp)	Minimum distance of markers (bp)
EG51_1	271	65,071,148	2,409,88.92	4,189,850	1
EG51_2	311	63,345,076	202,765.00	5,820,505	1
EG51_3	318	58,158,439	182,741.28	2,511,138	1
EG51_4	257	42,716,717	163,396.59	7,354,398	1
EG51_5	222	55,995,026	250,540.60	4,155,834	1
EG51_6	154	43,622,229	282,049.04	5,714,930	1
EG51_7	219	51,181,318	232,528.43	3,220,718	1
EG51_8	162	31,376,194	194,283.67	1,759,800	1
EG51_9	79	21,017,043	269,303.63	5,020,104	1
EG51_10	154	39,935,972	260,564.44	2,279,224	1
EG51_11	133	28,384,088	198,092.74	2,810,164	1
EG51_12	132	30,035,350	192,305.95	4,702,868	1
EG51_13	90	37,835,912	418,385.07	4,660,806	1
EG51_14	122	23,067,684	187,621.79	2,011,138	1
EG51_15	66	25,884,061	393,063.51	3,063,683	1
EG51_16	92	23,929,541	237,246.78	1,759,080	1
Sum	2,782	641,555,798			
Mean	173.87	40,097,237.38			

*EG: *Elaeis guineensis*; bp: base pair

III.1.2.4. Comparison between EG5.1 and PMv6 genome sequences

Data in Table IV shows the total SNP composition of the SNP physical positions on the reference genome of Singh *et al.*, (2013) (Eg5.1) and its improved version (i.e., EgPMv6) (Ong *et al.*, 2020).

This result was obtained after repositioning the two-reference genome to get more information and give a detailed explanation of the relationship between the two-oil palm reference genome. The total percentage of SNP composition in Table IV showed that there was no difference in the composition of SNP between the old references genome (Eg5.1) with the newly released version of the reference genome (i.e., EgPMv6) except for the case of chromosome 1 resulted in 85% of SNP composition with a total 797 SNPs in the chromosome

Eg51_1 on the old genome and 937 SNPs on chromosome GK000076.1 in the new EgPMv6 genome. Generally, we conclude the percentage of common SNPs between the two oil palm genomes is nearly 100% and no difference in the SNPs composition.

Table IV. Percentage of SNPs marker comparison between EG5.1 and PMv6 genome sequences.

Chromosome _Eg51 genome	Marker Number	Chromosome _Egpmv6 genome	Marker Number	SNPs on the Genome (%)
EG51_1	797	GK000076.1	937	85.06
EG51_2	700	GK000077.1	700	100.00
EG51_3	574	GK000078.1	578	99.31
EG51_4	494	GK000079.1	496	99.60
EG51_5	443	GK000080.1	443	100.00
EG51_6	309	GK000081.1	309	100.00
EG51_7	478	GK000082.1	478	100.00
EG51_8	328	GK000083.1	328	100.00
EG51_9	241	GK000084.1	241	100.00
EG51_10	394	GK000085.1	394	100.00
EG51_11	265	GK000086.1	265	100.00
EG51_12	295	GK000087.1	295	100.00
EG51_13	193	GK000088.1	193	100.00
EG51_14	323	GK000089.1	324	99.69
EG51_15	170	GK000090.1	170	100.00
EG51_16	199	GK000091.1	199	100.00
	6,203		6,350	

*EG: *Elaeis guineensis*, Egpmv6: *Elaeis guineensis palm modified version 6*

The newly released oil palm reference genome (i.e, EgPMv6) was also plotted against the old oil palm reference genome (i.e, Eg5.1) in the megabase pair (Fig. 26). The MareyMap plot showed almost the same in all the chromosomes found that, although some upturns existed (in particular for the smallest chromosomes), the positions on the two genomes are in general agreement and not that much difference between the two oil palm reference genomes in their arrangement.

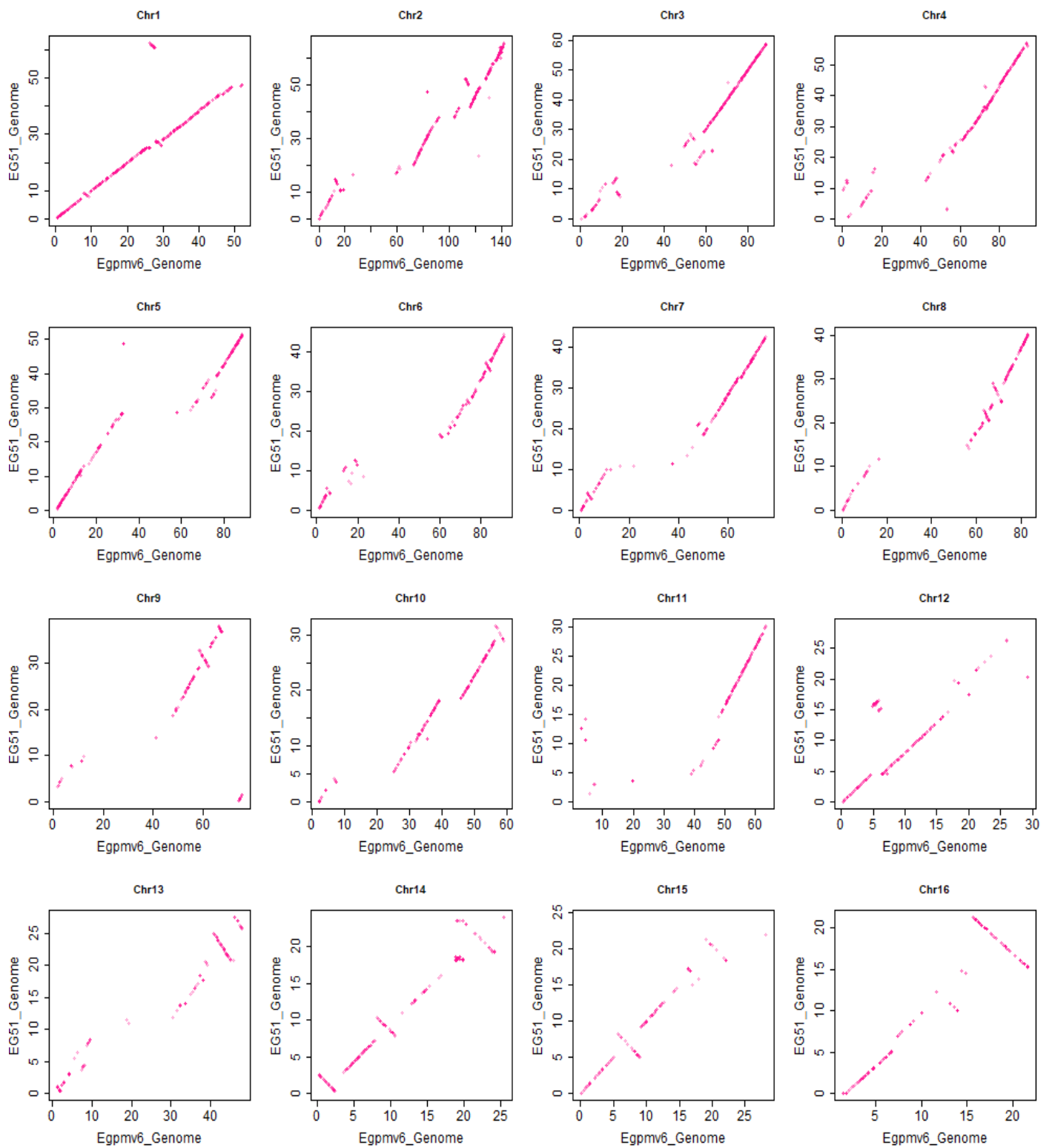


Fig. 26 Comparison of the SNP physical positions on the reference genome of Singh *et al.*, (2013) (Eg5.1) and its improved version Ong *et al.*, (2020) (EgPMv6). Distances are expressed in Mbp. *Chr: chromosome

III.1.3. Haplotype sharing between Deli and La Mé

The percentage of shared haplotypes between the Deli and La Mé oil palm breeding populations according to the length of the genomic window is represented in bp (Fig. 27) and cM (Fig. 28).

A large proportion of haplotypes were common between pairs of populations when considering short distances. Thus, 50% of the haplotypes with lengths around 30 bp (Fig. 27) and 40% of the haplotypes with lengths around 3,600 bp were common to the two populations, and 40% of the haplotypes with lengths around 0.20 cM were common to the two populations (Fig. 28). As expected when the length of the haplotypes increased, the percentage of shared haplotypes between populations decreased. The decrease was fast, with the percentage of common haplotypes falling below 20% for haplotypes longer than 300 kbp and 2.5 cM.

The frequency of the common haplotypes coincided to some extent for short haplotypes, while the differences increased for longer haplotypes. Thus, among the common haplotypes identified with a window size of 100 bp, more than one-half (51.6%) of the ones with a frequency >90% in Deli also had a frequency >90% in La Mé. This value fell to 25% for haplotypes identified with a window size of 50 kbp and to 14% for a window size of 500 kbp.

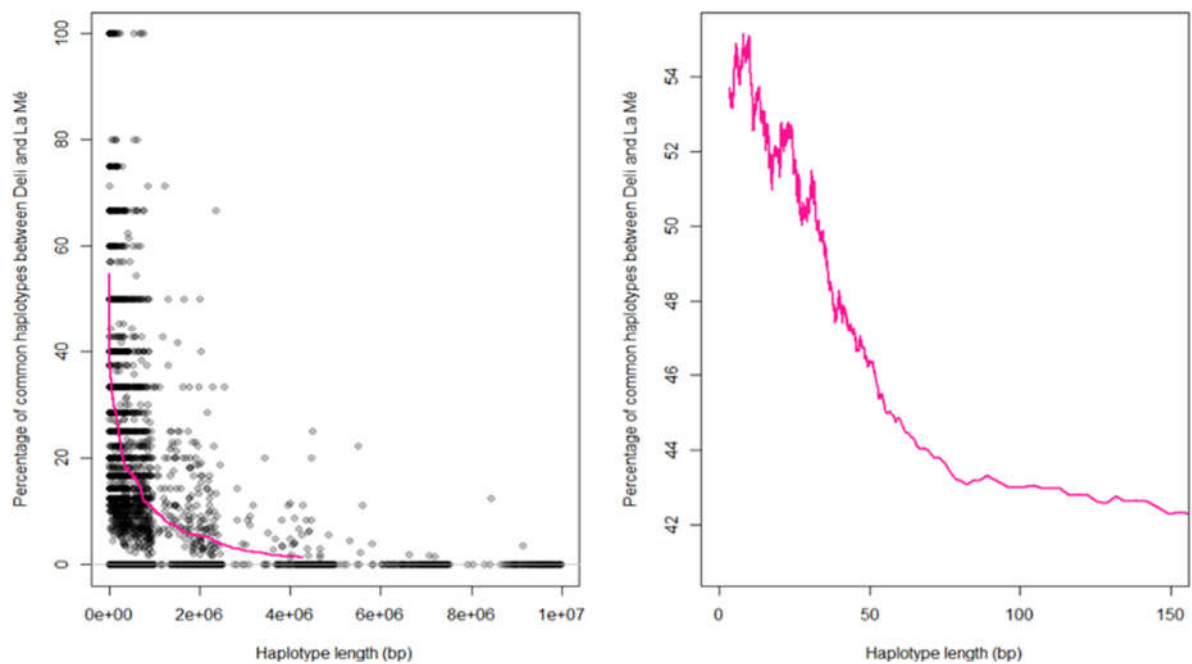


Fig. 27. Percentage of common haplotypes between Deli and La Mé oil palm breeding populations according to the haplotype length in bp. Each dot represents a haplotype. Color intensity indicates the density of overlapping dots. The smoothing curve in turquoise is the rolling average.

**bp: base pair*

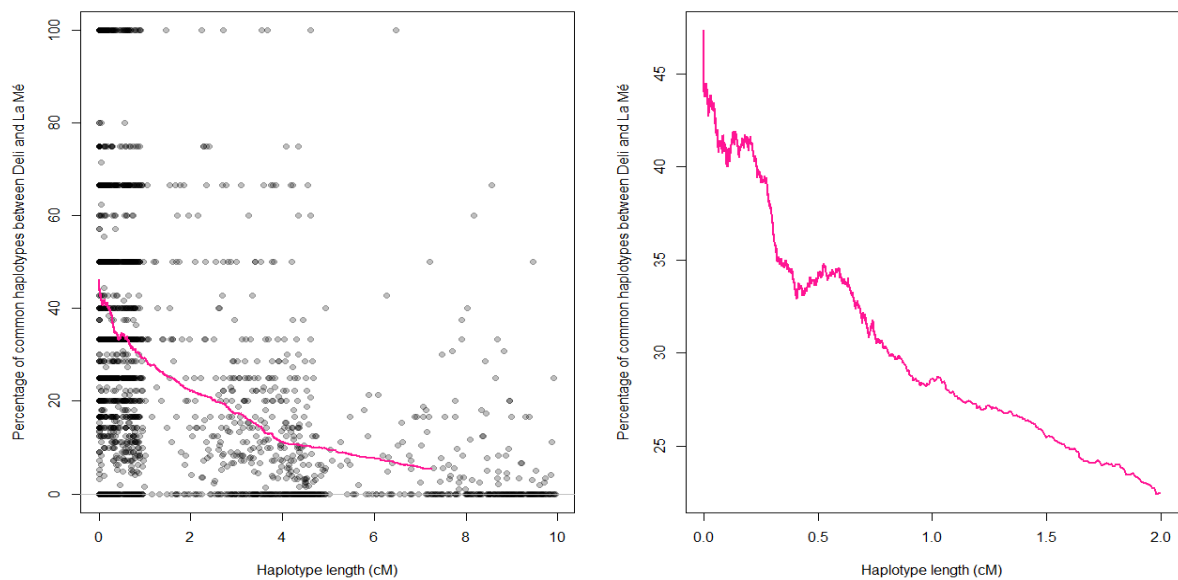


Fig. 28. Percentage of common haplotypes between Deli and La Mé oil palm breeding populations according to the haplotype length in cM. Each dot represents a haplotype. Color intensity indicates the density of overlapping dots. The smoothing curve in turquoise is the rolling average.

**cM: centiMorgan*

III.1.4. Effective size between Deli and La Mé

For both the Deli and La Mé oil palm breeding populations, multi-locus samples from a total of 7,324 SNP markers were used to calculate the population's effective size. The multi-locus samples were taken and run simultaneously for each breeding population and to avoid bias during computation we gave multi-locus samples having a code from both Deli and La Mé breeding populations.

Based upon the finding of the current study the lowest effective population size was observed from the Deli breeding population than the La Mé breeding population. Thus, the Deli breeding population with an effective population size value of 3.0. Whereas the largest effective population size was obtained from the La Mé breeding population with effective population size values of 3.6.

The Deli breeding population also has the lowest confidence intervals (CIs) value than the La Mé breeding population. Accordingly, a result pertaining Deli breeding population has a 2.7-3.3 CIs value at a 95% confidence interval. While the La Mé breeding population has a 3.0-5.2 CIs value at a 95% confidence interval.

III. 2. DISCUSSION

The contemporary thesis work was conducted to characterize the genome properties of Deli and La Mé oil palm (*Elaeis guineensis* Jacq.) breeding populations using the high throughput SNP marker obtained by genotyping by sequencing (GBS) method, which helps to identify polymorphisms distributed across the genome. To undertake these, major circumstances were considered: i.e, we evaluated some key genome properties factors that help to optimize the accuracy of GS in oil palm, and to do so we considered the major factors like LD, haplotype sharing, effective population size, fixation index, minor allelic frequency, heterozygosity and construction of genetic linkage map based on the Deli and La Mé oil palm breeding populations.

III.2.1. Genetic differentiation between Deli and La Mé

III.2.1.1. Distribution of minor allele and genotype frequencies across the population

The distribution of MAF revealed the majority of individuals in the Deli showed lower MAF than La Mé individuals which indicated that the Deli is distantly related to La Mé. This also resulted from their genetic history with more generations of selection, drift, and inbreeding in Deli than in La Mé. These results are consistent with the values reported by Cros *et al.* (2017) reported that a higher mean value of MAF was obtained from Group A, i.e, Deli with the MAF value of 0.15 than Group B, i.e, La Mé with a mean MAF value of 0.16. Again, Nyouma *et al.* (2020) also illustrated a higher mean value of MAF was obtained from La Mé populations (0.1) and a lower mean MAF value was obtained from Deli populations (0.07). By the same token, the Deli breeding population passes more generations of breeding activities like continuous selection, inbreeding, genetic drift, and narrow genetic base (less founder) than La Mé and resulting in a higher MAF.

This result also agrees with the work of Technow *et al.* (2014); Kumar *et al.* (2022) in maize and red clover Ergon *et al.* (2019) reported that lower MAF was observed from a breeding population that underwent more generations of continuous artificial selection than a breeding population that was not undergone this process.

III.2.1.2. Heterozygosity

The result showed that compared to the African sources of oil palm i.e., La Mé individual, less heterozygosity percentage was observed for Southeast Asia origin oil palm i.e., Deli individual. This is due to the Deli individual passing a lot of breeding activities like

continuous selection, inbreeding, genetic drift, and narrow genetic base (less founder) resulting in a more or less homozygote population than the African source oil palm. This result is in line with the work of Nyouma *et al.* (2020), who found lower genetic diversity was observed for the Deli than La Mé breeding populations. Likewise, Ajambang *et al.* (2016) reported that higher genetic variability was observed in African source oil palm than in Southeast Asia sources of oil palm. Conversely, Hayati *et al.* (2004) reported that higher genetic variability was observed in oil palm natural populations, i.e., oil palm from African sources than the Southeast Asian type of oil palm. Also, Rajanaidu *et al.* (2006) reported a higher rate of genetic variability was observed from the collected oil palm, i.e., African source than the standard variety, i.e., Deli dura due to the standard variety losing 36 alleles up on continuous selection. Harmoniously, Arias *et al.* (2012) pointed out higher genetic similarity was obtained from Group A oil palm (~76%) than the Group B oil palm with ~66% genetic similarity. Generally, we conclude that the four founder of Deli population were collected in Central Africa rather than in WA (Cochard *et al.*, 2009).

Further, the correlation of heterozygosity per SNPs between Deli and La Mé populations showed that the majority of SNPs were fixed on either side of the x and y-axis indicating there is higher heterozygosity between the two breeding populations. Similarly, the correlation in the frequency of alternate alleles per SNP between Deli and La Mé populations also showed there was a higher alternate alleles fixation alongside the x and y axes indicating the two breeding populations are distantly related. This fact is supported by scientific literature, for instance, historical evidence, agroecological fossil fuel, and physical evidence showed that generally, Africa is the center of origin of oil palm, and the Southeast Asia oil palm Deli dura material originated from an unknown area of Africa planted in Indonesia in 1848 (Corley & Tinker, 2016; Ithnin & Din, 2020). Furthermore, there is an indication Deli population is more genetically coherent with the Angola oil palm population than the WA source oil palm population (Campos & Caligari, 2017). Equally, Hayati *et al.* (2004) studied the genetic differentiation among African source oil palm populations and concluded that compared to the WA oil palm the Central and East Africa zone oil palm are genetically similar to each other this indicates the Deli oil palm is genetically closer to central Africa origin oil palm than WA type. Concerning this, there were known and unknown genetic materials transferred between the two geographical zones (i.e., Asia with Africa) a long time ago (Hartley, 1988; Hayati *et al.*, 2004; Cochard *et al.*, 2009).

Likewise, Cochard *et al.* (2009) illustrated the genetic origin of oil palm by using 18 individuals from 26 origins, and eight countries were analyzed with 14 microsatellite loci. They grouped the source of oil palm as African species at the Dahomey Gap, WA (Group I), “Benin-Nigeria-Cameroon-Congo-Angola” (Group II), and Deli group (Group III). From the results, they put an inference that the Deli group (Group III), derived from Group II, is due to the result of artificial selection (mass selection) and concluded that the four founders of the Deli population were collected in Central Africa rather than in WA. Equally important, Arias *et al.* (2013) studied morpho-agronomic and molecular characterization of oil palm using SSR markers and they concluded that there is a common morphological character shared by both Central African sources (i.e, Angola) and Asian (i.e, Deli) oil palm populations than the WA type oil palm.

III.2.1.3. Fixation index (F_{st})

The overall result of the fixation index showed that there is a significant degree of differentiation among the two pairs of oil palm breeding populations. The fixation index value between Deli and La Mé revealed that there is a significant degree of differentiation between the two oil palm breeding populations. On the other hand, the genetic differentiation among the oil palm populations could be a function of several factors, for instance, gene flow and natural or artificial selection, migration, and genetic drift (Hayati *et al.*, 2004; Gondro *et al.*, 2013). This result is also at par with the work of Corley & Tinker (2016); Campos & Caligari (2017); Soh *et al.* (2017) revealed that there was a less genetic relationship between Asians and WA oil palm populations. The fixation index obtained from this study between Deli and La Mé ($F_{st}=0.53$) was almost the same as the work of Cros *et al.* (2018) with the F_{st} of 0.55, using the same parental population and SNPs marker used and with the same population used in our fixation index result in almost 12.2% less than the previous result of (Cros *et al.*, 2015a). This is mainly because in this study we used higher SNPs data with 5% missing data and the same individuals that were used by Cros *et al.* (2017), then the simulation study of the same individuals with SSR data with missing data (Cros *et al.*, 2015a).

Corrospindgly, Jin *et al.* (2016) conducted the genetic differentiation between oil palm populations, and from their result, they concluded that there was higher genetic differentiation among palms from different geographical regions, and lower variation among Southeast Asian Dura and *Pisifera* palms than in Africa origin oil palm. Similarly, Hayati *et al.* (2004) compared the F_{st} value among oil palm populations originating from a different geographic region in

Africa, and they got $F_{st}=0.256$ in WA (i.e, Senegal, Guinea, Sierra Leone, Ghana, Nigeria, and Cameroon), Central Africa zone (The Democratic Republic of the Congo, Angola, and Tanzania) with the F_{st} of 0.073 and East Africa zone (Madagascar) with the F_{st} of 0.055. They concluded that oil palm populations from WA are genetically different from Central and East Africa. Our results are also in agreement with the oil palm study of Cochard *et al.* (2009), who concluded that the Deli population derived from a group comprising Benin, Nigeria, Cameroon, Congo, and Angola populations, while the populations west of Benin were genetically more different from Deli. This supports the idea that the four founders of the Deli population were collected in Central Africa rather than in WA (Cochard *et al.*, 2009).

The variation found in the F_{st} profile, which reached high values (>0.6) in some genomic regions, suggests that the fixation index is likely to be of interest in studying signatures of selection. This work at par with the work of Ye *et al.* (2020) reported that the F_{st} profile helps to reveal genomic regions of selective sweeps and is useful for locating SNPs under selection pressure (Chang *et al.*, 2018). This could help identify candidate genes, especially for traits with contrasting phenotypic values between breeding populations, such as BN and BW between A and B groups. However, a higher SNP density seems necessary to obtain clearer profiles with more pronounced peaks Porto-Neto *et al.* (2013), which could be linked to genes of interest based on the available information on oil palm annotation.

III.2.2. Within-population linkage disequilibrium and persistence phase between Deli and La Mé populations

III.2.2.1. Within-population linkage disequilibrium

The pattern of LD is one of the utmost factors affecting both GWAS and GS since both methods rely on LD between markers and causal polymorphisms (Sorkheh *et al.*, 2008; Hayes *et al.*, 2009; Yadav *et al.*, 2021). LD is thus one of the major factors that determine the number of markers required (Heffner *et al.*, 2009; Lebedev *et al.*, 2020). r^2 values of 0.3 are considered a minimum to get reliable results in GS and GWAS studies (Bejarano *et al.*, 2018). Here, when considering the genetic distances, the r^2 value reached 0.3 with SNPs separated by around 1.05 cM in Deli and 0.9 cM in La Mé. As our genetic map spanned 1,778.52 cM, achieving this distance between adjacent SNPs requires around 1,700 SNPs for Deli and 2,000 SNPs for La Mé. When considering the physical distances, the r^2 value of 0.3 was achieved with SNPs separated by around 220 kbp (0.22Mbp) in Deli and 210 kbp (0.21Mbp) in La Mé and here the genome length covered by SNPs spanned 643 Mbp, achieving these distances between adjacent

SNPs would take around 2,900 SNPs in Deli and 3,100 SNPs in La Mé, which can be considered close to the value obtained from the LD decay along with the genetic map. Considering that the goal should be to cover the whole genome and that the oil palm genome spans 1.8 gigabases (Singh *et al.*, 2013), 10,000 SNPs would be enough to reach the r^2 value of 0.3 in the two populations studied here (as this corresponds to around 8,200 SNPs in Deli and 8,600 La Mé). The effect of marker density on GS accuracy has already been evaluated on oil palm datasets comprising the populations considered here. It showed that depending on the study and trait, the number of SNPs required to achieve maximum GS accuracy was found to range from 500 to 7,000 (Cros *et al.*, 2017; Nyouma *et al.*, 2020). This is in agreement with the results obtained from the LD analysis.

Our results also revealed that the speed and the magnitude of LD decay varied between the breeding populations. In all the genomic regions the fastest LD decay was observed for the Deli population. While the lowest LD decay was observed by the La Mé population. The fastest LD decay over all the genomic regions and in both physical and genetic distance at Deli might be the history of the population, for instance, continuous selection of this population for breeding purposes resulted in less genetic diversity and a higher rate of gene flow within the population and create more or less the same populations (de Roos *et al.*, 2008). By the same token, the two populations were submitted to a founding bottleneck of similar magnitude. A bottleneck increases LD and slows down the LD decline (Tenaillon *et al.*, 2008). We can assume the higher value of LD in the Deli population in all genomic distances resulted from the fact that its history was marked by a larger number of generations marked by selection and inbreeding than in La Mé, with the bottleneck event in the Deli history dating back to 1848 against the 1920s in La Mé. Again, the rapid decline in the average r^2 of Deli compared to the decrease of r^2 in La Mé can be associated with differences in the effective population size of the breeds (Biegelmeyer *et al.*, 2016). On the other hand, the highest LD levels observed in Deli over all the genomic regions compared to La Mé can be related to one or more of the following factors: a higher ancestral relatedness, a historically smaller effective population size (founder effect), or a recent population reduction due to a bottleneck event and genetic drift, which probably occurred in Deli breeding population than La Mé (Bejarano *et al.*, 2018).

III.2.2.2. Persistence phase between Deli and La Mé populations

Based on the degree of genetic difference between breeds, the persistence of the allele phase was utilized to estimate their history and genetic ties (de Roos *et al.*, 2008). With

increasing marker distance, the correlation of r between breeding populations fell across all populations. High correlation of r values between populations ($r_{LD} > 0.6$, corresponding to $r_{LD}^2 > 0.25$) were obtained considering the markers that were the closest from each other, i.e. with distances < 0.5 cM on the genetic map or < 1 kbp on the physical map. Similarly, a large proportion of haplotypes was common between Deli and La Mé when considering windows of reduced size, with $> 40\%$ of haplotypes with lengths below 3,600 bp or 0.20 cM being common in the two populations. This explains the results of Nyouma *et al.* (2022, 2020), who found, using the same breeding populations and the same genotyping approach (GBS), that for GS predictions in oil palm, it was better not to model the parental origin of marker alleles. The superiority of GS models ignoring the parental origin of marker alleles over models considering it does not imply a complete persistence of phases between markers and QTLs among populations. Indeed, models that consider the parental origin of marker alleles are more complex and require the estimation of more parameters, possibly reducing their predictive ability, despite their ability to better depict the genetic differences between the population. The current study and the previous results of Nyouma *et al.* (2022, 2020), indicate that the level of conservation of phases among the Deli and La Mé populations captured with the present marker density is high enough to favor models ignoring the parental origin of marker alleles. A similar conclusion was reached by Technow *et al.* (2012) in maize, to explain the cases where this type of model outperformed the population-specific allele models. To further investigate this aspect, it would be interesting to study the correlation of marker effects obtained by GS models between Deli and La Mé populations, as done for maize (Technow *et al.*, 2014). To our knowledge, this is the first study investigating the persistence of LD and phases between oil palm populations. In accordance with this, the variation in the persistence phase over close markers could also be a result of a high degree of differentiation due to reproductive isolation, no recent common founder effect, different artificial selection between Deli than in La Mé and the four founder of Deli population were collected in Central Africa rather than in WA (Cochard *et al.*, 2009).

Other studies investigated the pattern of LD in oil palm, in particular Kwong *et al.* (2016) and Teh *et al.* (2016), using high-density SNP arrays. However, the results are difficult to compare, as the studies involved different populations, in particular inter-group hybrids, against parental populations in our study. However, Kwong *et al.* (2016) included in their work two breeding populations, JL \times DA and GM \times DA, that were mostly of Deli origin. Their LD value decreased by 50% from around 25 kbp to 200 kbp, i.e. in the same range as the value

found in our study (around 175 kbp). A previous study considered the same breeding populations as in the present study but used SSR markers (Cochard, 2008). The results were however in agreement, with Deli having the highest LD values. The consistency of these results shows that GBS is a suitable approach for LD studies, despite a higher rate of missing values and genotyping errors compared to SNP arrays and SSR, while providing much higher marker density than SSRs.

III.2.2.3. Comparison of genetic and physical maps

In oil palm, for the past 20 years, many genetic linkage maps have been constructed. The first genetic linkage map was constructed using RFLP markers (Mayes *et al.*, 1997). Since then, both dominant and co-dominant molecular markers have been used for the construction of genetic linkage maps. The construction of a genetic linkage map using SNP markers is now common in oil palms (see, for instance, (Jeenor & Volkaert, 2014; Ting *et al.*, 2014; Lee *et al.*, 2015; Pootakham *et al.*, 2015; Bai *et al.*, 2018a, 2018b; Gan *et al.*, 2018; Ong *et al.*, 2019; Herrero *et al.*, 2020; Ong *et al.*, 2020)).

Overall, the genetic linkage maps helped to identify genomic regions having major genes and quantitative trait loci (QTLs) that control oil yield Montoya *et al.* (2013); Jeenor & Volkaert (2014); Tisné *et al.* (2015), quality traits Singh *et al.* (2009); Pootakham *et al.* ((2015); Ong *et al.* (2019), vegetative growth Ukoskit *et al.* (2014); Lee *et al.* (2015); Bai *et al.* (2018b) and resistance to diseases (Tisné *et al.*, 2017; Daval *et al.*, 2021). High-density maps were also used to improve the assembly of previously published genome sequences by assigning scaffolds originally unplaced (Ong *et al.*, 2020). To our knowledge, the present study involved the largest number of individuals genotyped for the construction of a genetic map in oil palms. Another original aspect of our genetic map is the use of complex plant material including several families with varying degrees of relatedness, several generations, and different populations. By contrast, the previously published oil palm genetic maps were usually constructed from full-sib families (e.g. Watson *et al.* (2001); Cochard *et al.* (2009); Ting *et al.* (2013); Ukoskit *et al.* (2014); Ong *et al.* (2020)), although Billotte *et al.* (2010) used a factorial design. To our knowledge, only Cochard *et al.* (2009) and Daval *et al.* (2021) constructed genetic maps from populations with similar levels of complexity. However, they used SSR markers and the CRI-MAP software Green *et al.* (1990), which cannot handle thousands of SNPs.

The linkage map constructed in the present study spanned a total length of 1,778.52 cM, which is higher than the length of previously published genetic maps in oil palm made with

SNPs markers and LepMap software. For example, Herrero *et al.* (2020) obtained a map spanning 1,370 cM using a Cameroon×Nigeria cross and SNPs from SPET, and Ong *et al.* (2019, 2020) obtained maps of 1,151.7 cM, 1,268.26 cM, and 1,646.95 cM for Deli×AVROS, Deli Johore Labis×Nigeria and Deli×Nigeria populations, respectively, genotyped with an SNP array. The map of our current study is shorter than the map of Cochard *et al.* (2015), which reached 1,935 cM and was obtained using a similar oil palm population, SSR markers, and CRI-MAP software (Green *et al.*, 1990). This might be a consequence of the marker type, as it was shown that SSRs led to inflated maps compared to SNPs (Ball *et al.*, 2010).

The linkage map presented here, with an average marker density of one SNP in every 0.67 cM when considering unique positions, had a denser genome coverage compared to most previously published SNPs oil palm genetic linkage maps, like Ting *et al.* (2014) with one marker in every 1.40 cM and Pootakham *et al.* (2015) with one marker in every 1.26 cM. However, our map is less dense than the genetic linkage maps constructed by Ong *et al.* (2019, 2020) with one marker in every 0.04 cM, 0.05 cM, and 0.18 cM, depending on the map, and Bai *et al.* (2018a) with one marker every 0.29 cM and Herrero *et al.* (2020) with one marker in every 0.57 cM. Most of these variations in terms of the marker density of the genetic maps can be explained by differences in genotyping approaches and the size of the populations (Ferreira *et al.*, 2006; Semagn *et al.*, 2006). Combining high throughput genotyping and populations with at least 150 individuals appears as an efficient strategy to maximize marker density, as in Ong *et al.* (2019, 2020); Bai *et al.* (2018a), and the present study.

There were several upturns between the genetic and physical maps. For example, LG 1, 2, 5, and 7 had large upturns for regions of the genome of more than 10 Mbp. Aside from potential genome assembly artifacts, this can be the consequence of genomic rearrangements between populations, as the reference genome was obtained on an individual of the AVROS oil palm population Singh *et al.* (2013), which thus differed from the populations used for the genetic mapping. This aspect deserves further investigation, which could be done using population-specific genetic maps and reference genomes. This requires new data, with more genotyped individuals per population and new reference genomes.

Further, the recombination rate uniformity can be measured by the recombination rate concerning physical distance (Reich *et al.*, 2001). The average recombination rate was estimated at 2.85 cM/Mb. This value is in agreement with the ones found by Ong *et al.*, (2020) considering the same reference genome as in our study, i.e. 1.75 cM/Mb, 2.50 cM/Mb, and

1.93 cM/Mb in Deli×AVROS, Deli×Nigeria, and Deli Johore Labis populations, respectively. Variations in recombination rate along the chromosomes were noted in some chromosomes. In some cases, e.g. in chromosomes EG51_5, EG51_6, EG51_9, EG51_10, and EG51_15, they led to sigmoidal curves, which are expected under the effect of a lower recombination rate in the centromeric region (Semagn *et al.*, 2006; Ong *et al.*, 2020). For other chromosomes, these variations led to segments with lower SNP density compared to the rest of the chromosome and that corresponded to centromeric regions identified by (Singh *et al.*, 2013). This was for example the case in the 15 to 20 Mbp region with lower marker density in chromosomes EG51_11 and EG51_12.

III.2.2.4. Comparison between EG5.1 and PMv6 genome sequences

The comparison between the position of our SNPs on Eg5.1 with their position on EgPMv6, a new version of Eg5.1 improved through the use of a high-density linkage map Ong *et al.* (2020) found that, although some upturns existed (in particular for the smallest chromosomes), the positions on the two genomes are in general agreement. For example, LG 2 had 100% of its SNP located on the same chromosome according to Eg5.1 and EgPMv6 (Eg5.1_1 and GK000077.1, respectively), and almost identical SNP order between the two assemblies. This suggests that there is no considerable difference in genome assembly and gene annotation in the gene content between the two references genome in all chromosomes and our finding disagreement with the work of Bayer *et al.* (2017) comparing two *Brassica napus* references genome assembly outlined that the variation between two reference genome in the gene content it is due to the variation in genome assembly and annotation.

The result also showed that there is not that much vertical gap between the two references genome in the majority chromosome indicating that no such clear variation between the two reference genomes. Nevertheless, Herrero *et al.* (2020) indicate a vertical gap mainly due to variation in the references genome, and Ong *et al.* (2020) mainly work on the improvement of the scaffold and each pseudomolecule in the PMv6 genome and assembled using 142 scaffolds with an average length of 73.7 Mb, whereas only 19 scaffolds giving an average length of 41.1 Mb were reported in Singh *et al.* (2013) and this result the improvement of scaffold assignment to oil palm pseudomolecules from 43% to 77% this may increase the number of Single Nucleotide Variants (SNPs) without changing the gene content over the genome and a similar result was reported by Pan *et al.* (2019) in the human genome and indicating the new reference genome helps to increase the number of Single Nucleotide

Variants (SNPs) without changing the gene composition in two different human reference genomes (HG19 and HG38). The result from Galla *et al.* (2018) showed that GBS-based SNPs discovery from closely related reference genome correlates more significantly than distantly related species and which supports our findings. Generally, by re-conducting the result using the two oil palm reference genomes (i.e, EG5.1 and PMv6) the result will be quite similar.

III.2.3. Haplotype sharing between Deli and La Mé

The lowest shared haplotype between the two breeding populations showed that oil palms from Deli and La Mé are distantly related to each other. The lowest level of shared haplotype observed between the two breeding populations could be the result of the history of the two breeding populations indeed the four founders of the Deli population collected in Central Africa rather than in West Africa with no recent common founder effect and also reproductive isolation with the aid of different artificial selection in Deli than in La Mé Hartley (1988); Hayati *et al.* (2004); Cochard *et al.* (2009), as noted in sheep (Kijas *et al.*, 2012). In the same vine, the lower shared haplotype between the two oil palm breeding populations could also be the result of a lack of a different set of ancestral gene contributions to the modern cultivated oil palm (i.e, Deli dura), a similar observation reported by Hufford *et al.* (2013) in maize.

In this study, we found that a significant amount of haplotypes were common between the two breeding populations when using window lengths up to 30 bp. The higher shared haplotype at the beginning of the genomic region in both distances could also be due to the two breeding populations having no recent common founder this result agrees with the work of Coffman *et al.* (2020) in maize and outlined that no recent common haplotype sharing between maize breeding populations in the small genomic region could be a result of having a common founder but, no recent common founder. For oil palm, this corresponds to using at least 10,000 SNPs marker and our result suggests that, with such a marker density, multi-population GS could be worthwhile. This result is better than previously published by Cros *et al.* (2017) recommended for better GS accuracy the use of 5000 SNPs with less than 5% missing helps to capture genetic differences within parental families and less than by Nyouma *et al.* (2020) by least 7000 SNPs with the secondary priority percentage of missing data per SNPs.

The haplotype sharing between populations is also one of the factors that affect the accuracy of genomic selection (Calus *et al.*, 2008; Bhat *et al.*, 2021). A report by Varshney *et al.* (2005); Jannink *et al.* (2010); Qian *et al.* (2017) outlined that the selection efficiency of both

MAS and GS improved by the arrangement of haplotype genes from different breeding. The subsequent identification and characterization of functionally significant genomic areas throughout evolution and/or selection can be done using a conserved haplotype structure (Rahman *et al.*, 2022). For example, *Eucalyptus* Ballesta *et al.* (2019) outlined that a model having a haplotype effect (either HAP or HAP-SNP) helps to increase the prediction accuracy of low-heritability traits, for instance, for stem straightness ($r=0.58$). Conversely, in soybean haplotype-based selection helps to select the genomic region that controls plant height (Bhat *et al.*, 2022). By the same token, a report from Zhao *et al.* (2022) on tomatoes showed that haplotype-based analysis helps to identify promising candidate genes that control tomato fruit weight and metabolite contents. In livestock, Goddard & Hayes (2007) also outlined that the accuracy of GS increases by estimating the haplotype effects as the amount of data (i.e., number of animals with phenotypes and marker genotypes) for estimation increases, especially at lower marker densities. Furthermore, a report from Wientjes *et al.* (2013) showed that compared to the individual shared haplotype length, genomic selection accuracy increases by the accumulated length of shared haplotypes between two individuals.

III.2.4. Effective size between Deli and La Mé

The size of an effective population is linked to the population's history Caballero (1994) and it's useful for evolutionary biology, conservation genetics, and plant and animal breeding due to the fact that it tracks the rates of genetic drift and inbreeding and influences the effectiveness of deterministic evolutionary forces like mutation, selection, and migration (Wang *et al.*, 2016).

In oil palm, till today there was no estimate available of effective population size for La Mé breeding populations. The small values obtained are not surprising given the history of the populations, with a small number of founders and under the effect of inbreeding. In Cros *et al.* (2014), effective population size was estimated for a subset of 104 Deli individuals from the population used here, with 16 SSR markers chosen on different LGs and the LD method (Waples & Do, 2008). This gave a N_e of 5 ± 1.1 (SD), i.e. similar to the result we obtained here. This indicates the robustness of the method against marker type and density. The smallest value for the Deli population obtained despite its larger number of founders may result from the fact, already mentioned above, that one of the founders had a much greater contribution than the other founders. The small effective population size values obtained here also explain the fact that GS can be implemented with small training populations and low marker density.

Thus, in previous studies, GS models trained with data from only 108 Deli and 102 La Mé individuals were efficient enough to replace phenotypic selection before clonal trials Nyouma *et al.* (2020) while GS accuracy plateaued with only 500 to 2,000 SNPs, depending on the trait (Cros *et al.*, 2017).

The lower effective population size in the Deli population over the La Mé population is also related to their difference in LD since, effective population size and LD have an inverse relationship, with high rates of genetic drift and inbreeding in low effective population size populations leading to strong LD between markers and QTLs compared to high effective population size populations (Grattapaglia, 2014; Lin *et al.*, 2014; Thistlethwaite *et al.*, 2020). In our cases, the Deli population has a small N_e size which results in a rapid decline in the average r^2 (Small LD) than the La Mé breeding population and a similar result was also reported by (Makina *et al.*, 2015; Biegelmeyer *et al.*, 2016; Bejarano *et al.*, 2018).

The result also revealed that a smaller effective population size by the Deli population is also an indication of losing genetic diversity more quickly than the La Mé breeding populations due to inbreeding and genetic drift which affects the capacity of released individuals to survive and reproduce in the wild. This condition has been linked to an increased risk of population extinction, a slowed rate of population growth, a diminished capacity to respond to environmental change, and a decreased capacity for disease resistance (Kliman *et al.*, 2008; Furlan *et al.*, 2012; Yates *et al.*, 2019). Accordingly, Wang *et al.* (2016) suggested that lower effective population size populations need different ecological, evolutionary, and conservation breeding and genetic approaches. Therefore, for future use of oil and its products from Deli breeding population we need to apply the above-mentioned approaches.

**CONCLUSION,
RECOMMENDATIONS AND
PERSPECTIVES**

CONCLUSION AND PERSPECTIVES

IV.1. CONCLUSION

Studies on oil palm Genomic Selection gave promising results, but the method could be optimized. For that purpose, the current study investigated the genome properties of two main oil palm breeding populations, Deli and La Mé, used in the RRS breeding scheme. Specifically, the study focused on minor allelic frequency, heterozygosity, LD, N_e , haplotype sharing, and F_{st} .

A high-density genetic map was constructed from a complex population including several families with varying sizes and levels of relatedness and with different genetic backgrounds. It included 4,252 SNPs from GBS and spanned 1,778.52 cM, with an average recombination rate of 2.85 cM/Mbp. The LD $r^2 = 0.3$ spanned over 1.05 cM/0.22 Mbp in Deli and 0.9 cM/0.21 Mbp in La Mé. When considering the genetic distance with $r^2 = 0.3$, 1,700 SNPs for Deli and 2,000 SNPs for La Mé were required whereas when considering the physical distance, around 2,900 SNPs in Deli and 3,100 SNPs in La Mé were required. Deli has the fastest LD decay over the genomic region in both physical and genetic distances. A high correlation of LD between populations ($r_{LD} > 0.6$) was obtained when considering the markers separated by short distances, i.e. < 0.5 cM on the genetic map or < 1 kbp on the physical map. The percentage of common haplotypes was above 40% for short haplotypes (3,600 bp or 0.20 cM). This resemblance decreased with the distance between SNPs, with for example the percentage of common haplotypes falling below 20% for haplotypes longer than 300 kbp. The F_{st} was high (0.53). In the two populations, 10,000 SNPs would be enough to reach this level of LD, which is advisable given the small N_e values of the current populations ($N_e < 5$). Overall, the results showed strong genetic differentiation between Deli and La Mé, and this was approved by F_{st} , correlation of heterozygosity per SNP, correlation of frequency of alternate allele per SNP, and percentage of common haplotypes between populations. The level of resemblance between them over short genomic distances likely explains the superiority of GS models ignoring the parental origin of marker alleles over models taking this information into account.

It is suggested that oil palm breeding programs be used to promote genetic gain from genomic selection in oil palms.

IV.2. RECOMMENDATIONS

- In both breeding populations, 10,000 SNPs marker density would be enough to reach this level of LD.
- Before applying GS studies in other oil palm breeding materials breeder should implement the genome properties study first.
- For the subsequent GS studies in oil palm for better genome accuracy, GS studies should take into account the genetic differentiation between Deli and La Mé.
- Enlarging the genetic diversity in Deli and La Mé for better genetic progress.

IV.3. PERSPECTIVES

The present study showed an interest in understanding the genome properties that directly and indirectly affect the accuracy of genomic prediction methods used for the genetic improvement of palm oil yield. However, to attain the world palm oil demand with the current climate change, population growth, and other biotic and abiotic stresses, future oil palm research should focus on:

- Population-specific genetic maps;
- Quantifying the study again using new reference genomes;
- Studying this with other genome properties factors;
- Studying this work with other populations like AVROS, Sibiti, Yangambi, etc
- Use of GS models ignoring the parental origin of marker alleles;
- Haplotype-based genomic prediction;

REFERENCES

REFERENCES

- Ajambang W., Ngando Ebongué G., Bakoumé C., Ataga C., Okoye M. N., Enaberue L., Etta C. E., Konan J. N., Allou D., Diabaté S. & Konan E., 2016. Oil Palm Breeding and Seed Production in Africa. In: International Seminar on Oil Palm Breeding and Seed Production. Kisana, Sumatra, Indonesia. 29: 1–30.
- Akdemir D. & Isidro-Sánchez J., 2019. Design of training populations for selective phenotyping in genomic prediction. *Scientific reports*, 9(1): 1–15.
- Arias D., González M., Prada F., Restrepo E. & Romero H., 2013. Morpho-agronomic and molecular characterisation of oil palm *Elaeis guineensis* Jacq. material from Angola. *Tree Genetics & Genomes*, 9(5): 1283–1294.
- Arias D., Montoya C., Rey L. & Romero H., 2012. Genetic similarity among commercial oil palm materials based on microsatellite markers. *Agronomía Colombiana*, 30(2): 188–195.
- Babu B.K. Mathur R. Anitha P., Ravichandran G. & Bhagya H., 2021. Phenomics, genomics of oil palm (*Elaeis guineensis* Jacq.): way forward for making sustainable and high yielding quality oil palm. *Physiology and Molecular Biology of Plants*, 27: 587–604.
- Babu B.K. & Mathu R. K., 2016. Molecular breeding in oil palm (*Elaeis guineensis*): Status and future perspectives. *Prog. Hort*, 48(2): 123–131.
- Bai B., Wang L., Zhang Y. J., Lee M., Rahmadsyah R., Alfiko Y., Ye B. Q., Purwantomo S., Suwanto A., Chua N.-H. & Yue G. H., 2018a. Developing genome-wide SNPs and constructing an ultrahigh-density linkage map in oil palm. *Scientific Reports*, 8(1): 1–7.
- Bai B., Zhang Y. J., Wang L., Lee M., Rahmadsyah Ye, B. Q., Alfiko Y., Purwantomo S., Suwanto A. & Yue G. H., 2018b. Mapping QTL for leaf area in oil palm using genotyping by sequencing. *Tree Genetics & Genomes*, 14(2): 1–9.
- Baker W. J., Norup M. V., Clarkson J. J., Couvreur T. L. P., Dowe J. L., Lewis C. E., Pintaud J.-C., Savolainen V., Wilmot T. & Chase M. W., 2011. Phylogenetic relationships among arecoid palms (Arecaceae: Arecoideae). *Annals of Botany*, 108(8): 1417–1432.
- Ball A. D., Stapley J., Dawson D. A., Birkhead T. R., Burke T. & Slate J., 2010. A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*). *BMC genomics*, 11(1): 1–15.
- Ballesta P., Maldonado C., Pérez-Rodríguez P. & Mora F., 2019. SNP and Haplotype-Based Genomic Selection of Quantitative Traits in *Eucalyptus globulus*. *Plants*, 8(9): 331.

- Barcelos E., Rios S. de A., Cunha R. N., Lopes R., Motoike S.Y., Babiychuk E., Skiryecz A. & Kushnir S., 2015. Oil palm natural diversity and the potential for yield improvement. *Frontiers in plant science*, 6:190.
- Bayer P.E., Hurgobin B., Golicz A.A., Chan C.K., Yuan Y., Lee H., Renton M., Meng J., Li R. & Long Y., 2017. Assembly and comparison of two closely related *Brassica napus* genomes. *Plant biotechnology journal*, 15(12): 1602–1610.
- Bayer P.E., Petereit J., Danilevicz M.F., Anderson R., Batley J. & Edwards D., 2021. The application of pangenomics and machine learning in genomic selection in plants. *The Plant Genome*, 14(13): 1-9.
- Beirnaert A.D.F. & Vanderweyen R., 1941. Contribution à l'étude genetique et biometrique des variétés d'*Elaeis Guineensis Jacquin*. Publ. Inst. Nat. Etude Agron. Congo Belge Ser. *Sci.*, 27: 1–101.
- Bejarano D., Martínez R., Manrique C., Parra L.M., Rocha J.F., Gómez Y., Abuabara Y. & Gallego J., 2018. Linkage disequilibrium levels and allele frequency distribution in Blanco Orejinegro and Romosinuano Creole cattle using medium density SNP chip data. *Genet Mol Biol*, 41: 426–433.
- Bernardo R.N., 2010. Breeding for quantitative traits in plants, 2nd ed. ed. Stemma Press, Woodbury, Minn. 369 p.
- Bevan M.W., Uauy C., Wulff B.B.H., Zhou J., Krasileva K. & Clark M.D., 2017. Genomic innovation for crop improvement. *Nature*, 543(7645): 346–354.
- Bhat J.A., Karikari B., Adeboye K.A., Ganie S.A., Barmukh R., Hu D., Varshney R.K. & Yu D., 2022. Identification of superior haplotypes in a diverse natural population for breeding desirable plant height in soybean. *Theoretical and Applied Genetics*, 135: 2407–2422.
- Bhat J.A., Yu D., Bohra A., Ganie S.A., & Varshney R.K., 2021. Features and applications of haplotypes in crop breeding. *Communications biology*, 4(1): 1–12.
- Biegelmeyer P., Gulias-Gomes C.C., Caetano A.R., Steibel J.P. & Cardoso F.F., 2016. Linkage disequilibrium, persistence of phase and effective population size estimates in Hereford and Braford cattle. *BMC Genetics*, 17(1): 1–12.
- Billotte N., Jourjon M.F., Marseillac N., Berger A., Flori A., Asmady H., Adon B., Singh R., Nouy B., Potier F., Cheah S.C., Rohde W., Ritter E., Courtois B., Charrier A. & Mangin B., 2010. QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis Jacq.*). *Theor Appl Genet*, 120(8):1673–1687.

- Billotte N., Marseillac N., Risterucci A.-M., Adon B., Brottier P., Baurens F.-C., Singh R., Herrán A., Asmady H., Billot C., Amblard P., Durand-Gasselín T., Courtois B., Asmonó D., Cheah S.C., Rohde W., Ritter E. & Charrier A., 2005. Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* Jacq.). *Theor Appl Genet*, 110(4): 754–765.
- Brandariz S.P. & Bernardo R., 2019. Small ad hoc versus large general training populations for genomewide selection in maize biparental crosses. *Theoretical and Applied Genetics*, 132(2): 347–353.
- Brauner P.C., Müller D., Molenaar W.S. & Melchinger A.E., 2020. Genomic prediction with multiple biparental families. *Theoretical and Applied Genetics*, 133(1): 133–147.
- Browning B.L., Zhou Y. & Browning S.R., 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics*, 103(3): 338–348.
- Caballero A., 1994. Developments in the prediction of effective population size. *Heredity*, 73(6): 657–679.
- CABI 2019. *Elaeis guineensis* (African oil palm). DistributionMaps <https://www.cabi.org/isc/datasheet/20295#toDistributionMaps>. Accessed January 20, 2021. <https://doi.org/10.1079/cabicompendium.20295>
- Cadena T., Prada F., Perea A. & Romero H.M., 2013. Lipase activity, mesocarp oil content, and iodine value in oil palm fruits of *Elaeis guineensis*, *Elaeis oleifera*, and the interspecific hybrid O×G (*E. oleifera* × *E. guineensis*). *Journal of the Science of Food and Agriculture*, 93(3): 674–680.
- Calabrese B., 2019. Linkage Disequilibrium, in: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), *Encyclopedia of Bioinformatics and Computational Biology*. Academic Press, Oxford, 763–765 p.
- Calleja-Rodríguez A., Pan J., Funda T., Chen Z., Baison J., Isik F., Abrahamsson S. & Wu H.X., 2020. Evaluation of the efficiency of genomic versus pedigree predictions for growth and wood quality traits in Scots pine. *BMC genomics*, 21(1): 1–17.
- Calus M.P.L., Meuwissen T.H.E., de Roos A.P.W. & Veerkamp R.F., 2008. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. *Genetics*, 178(1): 553–561.
- Campos G. de los, Hickey J.M., Pong-Wong R., Daetwyler H.D. & Calus M.P.L., 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics*, 193(2): 327–345.

- Campos H. & Caligari P.D.S., 2017. Genetic Improvement of Tropical Crops. Springer International Publishing, Cham. 320 p.
- 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 Science*, 284: 9–15.
- Cericola F., Lenk I., Fè D., Byrne S., Jensen C.S., Pedersen M.G., Asp T., Jensen J. & Janss L., 2018. Optimized use of low-depth genotyping-by-sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (*Lolium perenne* L.). *Frontiers in plant science*, 9: 369.
- Chang L.-Y., Toghiani S., Hay E.H., Aggrey S.E. & Rekaya R., 2019. A Weighted Genomic Relationship Matrix Based on Fixation Index (FST) Prioritized SNPs for Genomic Selection. *Genes (Basel)*, 10(11): 922.
- Chang L.-Y., Toghiani S., Ling A., Aggrey S.E. & Rekaya R., 2018. High density marker panels, SNPs prioritizing and accuracy of genomic selection. *BMC genetics*, 19(1): 1–10.
- Chen B.K., Seligman B., Farquhar J.W. & Goldhaber-Fiebert J.D., 2011. Multi-Country analysis of palm oil consumption and cardiovascular disease mortality for countries at different stages of economic development: 1980-1997. *Globalization and health*, 7:1–10.
- Clark S.A., Hickey J.M., Daetwyler H.D. & van der Werf J.H., 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics Selection Evolution*, 44:1–9.
- 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., Rekima S., Billotte N., de Chenon R.D., Koutou A., Nouy B., Omoré A., Purba A.R., Glazsmann J.-C. & Noyer J.-L., 2009. Geographic and genetic structure of African oil palm diversity suggests new approaches to breeding. *Tree Genetics & Genomes*, 5(3): 493–504.
- Cochard B., Amblard P. & Durand-Gasselín T., 2005. Oil palm genetic improvement and sustainable development. *OCL*, 12(2): 141–147.

- Cochard B., Carrasco-Lacombe C., Pomies V., Dufayard J.-F., Suryana E., Omoré A., Tristan D.-G. & Tisné S., 2015. Pedigree-based linkage map in two genetic groups of oil palm. *Tree Genetics & Genomes*, 11(4):1–12.
- Coffman S.M., Hufford M.B., Andorf C.M. & Lübberstedt T., 2020. Haplotype structure in commercial maize breeding programs in relation to key founder lines. *Theoretical and Applied Genetics*, 133(2): 547–561.
- Collard B.C.Y. & Mackill D.J., 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci.* 363(1491): 557–572.
- Collins A.R. (Ed.), 2007. Linkage Disequilibrium and Association Mapping: Analysis and Applications, *Methods in Molecular Biology*. Humana Press. 376 p.
- Combs E. & Bernardo R., 2013. Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *The Plant Genome*, 6(1): 1–7.
- Comstock R.E., Robinson H.F. & Harvey P.H., 1949. A Breeding Procedure Designed To Make Maximum Use of Both General and Specific Combining Ability¹. *Agronomy Journal*, 41: 360–367.
- Cook S., Whelan M.J., Evans C.D., Gauci V., Peacock M., Garnett M.H., Kho L.K., Teh Y.A. & Page S.E., 2018. Fluvial organic carbon fluxes from oil palm plantations on tropical peatland. *Biogeosciences*, 15: 7435–7450.
- Corbin L.J., Liu A., Bishop S. & Woolliams J., 2012. Estimation of historical effective population size using linkage disequilibria with marker data. *Journal of Animal Breeding and Genetics*, 129(4): 257–270.
- Corley R.H.V., 2009. How much palm oil do we need? *Environmental Science & Policy*, 12(2): 134–139.
- Corley R.H.V. & Tinker P.B., 2016. *The oil palm*, John Wiley-Blackwell & Sons, Ltd, Chichester-West Sussex UK, 700 p.
- Corley R.H.V. & Tinker P.B., 2015. *The oil palm*. John Wiley-Blackwell & Sons, Ltd, Chichester-West Sussex UK, 341 p.
- Corley R.H.V. & Tinker P.B.H., 2003. The Oil Palm, *The oil palm*. John Wiley-Blackwell & Sons, Ltd, Chichester-West Sussex UK, 141 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): 1–17.
- 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. & Clement-Demange A., 2019. Within-family genomic selection in rubber tree (*Hevea brasiliensis*) increases genetic gain for rubber production. *Industrial Crops and Products*, 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 Breeding*, 38(7): 89–101.
- Crossa J., Pérez-Rodríguez P., Cuevas J., Montesinos-López O., Jarquín D., de los Campos G., Burgueño J., González-Camacho J.M., Pérez-Elizalde S., Beyene Y., Dreisigacker S., Singh R., Zhang X., Gowda M., Roorkiwal M., Rutkoski J. & Varshney R.K., 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, 22(11): 961–975.
- Daetwyler H.D., Calus M.P.L., Pong-Wong R., Campos G. de los. & Hickey J.M., 2013. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics*, 193(2): 347–365.
- 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., McVean G. & Durbin R., 1000 Genomes Project Analysis Group, 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156–2158.
- Daval A., Pomiès V., Le Squin S., Denis M., Riou V., Breton F., Bink M., Cochard B., Jacob F. & Billotte N., 2021. In silico mapping in an oil palm breeding program reveals a

- quantitative and complex genetic resistance to *Ganoderma boninense*. *Molecular Breeding*, 41(9): 1–18.
- de Moraes B.F.X., dos Santos R.F., de Lima B.M., Aguiar A.M., Missiaggia A.A., da Costa Dias D., Rezende G.D.P.S., Gonçalves F.M.A., Acosta J.J. & Kirst M., 2018. Genomic selection prediction models comparing sequence capture and SNP array genotyping methods. *Molecular Breeding*, 38(9): 1–14.
- de Roos A.P.W., Hayes B.J., Spelman R.J. & Goddard M.E., 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics*, 179(3): 1503–1512.
- de Souza L.M., dos Santos L.H.B., Rosa J.R.B.F., da Silva C.C., Mantello C.C., Conson A.R.O., Scaloppi E.J.J., Fialho J de F., de Moraes M.L.T., Gonçalves P de S., Margarido G.R.A., Garcia A.A.F., Le Guen V. & de Souza A.P., 2018. Linkage Disequilibrium and Population Structure in Wild and Cultivated Populations of Rubber Tree (*Hevea brasiliensis*). *Front. Plant Sci*, 9: 815.
- Denis M. & Bouvet J.-M., 2013. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. *Tree Genetics & Genomes*, 9(1): 37–51.
- Dislich C., Keyel A.C., Salecker J., Kisel Y., Meyer K.M., Auliya M., Barnes A.D., Corre M.D., Darras K. & Faust H., 2017. A review of the ecosystem functions in oil palm plantations, using forests as a reference system. *Biological Reviews*, 92:1539–1569.
- Do C., Waples R.S., Peel D., Macbeth G.M., Tillett B.J. & Ovenden J.R., 2014. NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. *Molecular Ecology Resources*, 14(1): 209–214.
- Dudley J.W., 1997. Quantitative Genetics and Plant Breeding, in: Sparks, D.L. (Ed.), *Advances in Agronomy*. Academic Press, 1–23 p.
- 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 Genetics & Genomes*, 13(4): 1–12.
- Durand-Gasselín T., Blangy L., Picasso C., Franqueville H. de, Breton F., Amblard P., Cochard B., Louise C. & Nouy B., 2010. Sélection du palmier à huile pour une huile de palme durable et responsabilité sociale. *OCL*, 17: 385–392.
- Durand-Gasselín T., De Franqueville H., Amblard P., Breton F., Jacquemard J.-C., Syaputra I., Cochard B., Louise C. & Nouy B., 2011. Breeding for sustainable palm oil. *International*

- Society for Oil Palm Breeders (ISOPB) and Malaysian Palm Oil Board (MPOB). 178 – 193
- Durand-Gasselín T., Kouame Kouame R., Cochard B., Adon B. & Amblard P., 2000. Diffusion variétale du palmier à huile (*Elaeis guineensis* Jacq.). *OCL*, 7(2): 207–214.
- Ergon Å., Skøt L., Sæther V.E. & Rognli O.A., 2019. Allele frequency changes provide evidence for selection and identification of candidate loci for survival in red clover (*Trifolium pratense* L.). *Frontiers in plant science*, 10: 718.
- Falconer D.S. & Mackay T.F.C., 1996. Introduction to Quantitative Genetics, Subsequent edition. ed. Benjamin-Cummings Pub Co, Harlow. Essex, 464 p.
- Fanelli Carvalho H., Galli G., Ventrone Ferrão L.F., Vieira Almeida Nonato J., Padilha L., Perez Maluf M., Ribeiro de Resende Jr M.F., Guerreiro Filho O. & Fritsche-Neto R., 2020. The effect of bienniality on genomic prediction of yield in arabica coffee. *Euphytica*, 216(6): 1–16.
- Ferrão L.F.V., Ferrão R.G., Ferrão M.A.G., Fonseca A., Carbonetto P., Stephens M. & Garcia A.A.F., 2019. Accurate genomic prediction of *Coffea canephora* in multiple environments using whole-genome statistical models. *Heredity*, 122(3): 261–275.
- Ferreira A., Silva M.F. da Silva L. da C. e & Cruz C.D., 2006. Estimating the effects of population size and type on the accuracy of genetic maps. *Genet. Mol. Biol*, 29: 187–192.
- Flint-Garcia S.A., Thornsberry J.M. & Buckler IV E.S., 2003. Structure of linkage disequilibrium in plants. *Annual review of plant biology*, 54: 357–374.
- Florence J., Cros D. & Cochard B., 2017. ISOPB 2017 Agrigenomics in the breeder’s toolbox: latest advances towards an optimal implementation of genomic selection in oil palm. *Kuala Lumpur: Malaysian palm oil board-International Society for Oil Palm Breeders*, 21 p.
- Furlan E., Stoklosa J., Griffiths J., Gust N., Ellis R., Huggins R.M. & Weeks A.R., 2012. Small population size and extremely low levels of genetic diversity in island populations of the platypus, *Ornithorhynchus anatinus*. *Ecology and evolution*, 2(4): 844–857.
- Gabriel S.B., 2002. The Structure of Haplotype Blocks in the Human Genome. *Science*, 296 (5576): 2225–2229.
- Galla S.J., Forsdick N.J., Brown L., Hoepfner M.P., Knapp M., Maloney R.F., Moraga R., Santure A.W. & Steeves T.E., 2018. Reference genomes from distantly related species can be used for discovery of single nucleotide polymorphisms to inform conservation management. *Genes*, 10(1): 9.

- Gallais A. & Poly J., 1990. Théorie de la sélection en amélioration des plantes. Masson, Paris. 588 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 Genetics*, 59(1): 23–34.
- Gascon J.P. & 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(2): 75–84.
- Gianola D. & Van Kaam J.B., 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178: 2289–2303.
- 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.
- Goddard M.E., Hayes B.J. & Meuwissen T.H.E., 2010. Genomic selection in livestock populations. *Genet Res*, 92(5-6):413–421.
- Godswill N.-N., Frank N.-E.G., Walter A.-N., Edson M.-Y.J., Kingsley T.-M., Arondel V., Martin B.J. & Emmanuel Y., 2016. Chapter 10 - Oil Palm, in: Gupta, S.K. (Ed.), *Breeding Oilseed Crops for Sustainable Production*. Academic Press, San Diego, 217–273 p.
- Gois I.B., Borém A., Cristofani-Yaly M., Resende M.D.V., Azevedo C., Bastianel M., Novelli V., & Machado M., 2016. Genome wide selection in Citrus breeding. *Genetics and Molecular Research*, 15(4): gmr15048863
- Gondro C., Werf J. van der. & Hayes B., 2013. *Genome-Wide Association Studies and Genomic Prediction, Methods in Molecular Biology*. Humana Press. 566 p.
- Goode E.L., 2011. Linkage Disequilibrium, in: Schwab, M. (Ed.), *Encyclopedia of Cancer*. Springer, Berlin, Heidelberg, 2043–2048 p.
- 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: Volume 1. Managing, Sequencing and Mining Genetic Resources*. Springer Netherlands, Dordrecht, 651–682 p.
- Grattapaglia D. & Resende M.D., 2011. Genomic selection in forest tree breeding. *Tree Genetics & Genomes*, 7(2): 241–255.

- Grattapaglia D., Silva-Junior O.B., Resende R.T., Cappa E.P., Müller B.S.F., 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.
- Green P., Falls K. & Crooks S., 1990. Documentation for CRI-MAP, version 2.4. Washington University School of Medicine, St. Louis.
- Gupta P.K., Rustgi S. & Kulwal P.L., 2005. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant Mol Biol*, 57(4): 461–485.
- Hamazaki K. & Iwata H., 2020. RAINBOW: Haplotype-based genome-wide association study using a novel SNP-set method. *PLoS computational biology*, 16(2): e1007663.
- Hardon J.J., 1969. Interspecific hybrids in the genus *Elaeis* II. vegetative growth and yield of F1 hybrids *E. guineensis* x *E. oleifera*. *Euphytica*, 18(3): 380–388.
- Hardon J.J. & Tan G.Y., 1969. Interspecific hybrids in the genus *Elaeis* I. crossability, cytogenetics and fertility of F1 hybrids of *E. guineensis* x *E. oleifera*. *Euphytica*, 18(3): 372–379.
- Hartley C.W.S., 1988. The oil palm (*Elaeis guineensis* Jacq.). Longman Scientific & Technical ; Wiley, Harlow, Essex, England; New York. 781 p.
- Hartley C.W.S., 1977. The Oil Palm (*Elaeis guineensis* Jacq.). Longman Scientific & Technical, New York, 800 p.
- Hayati A., Wickneswari R., Maizura I. & Rajanaidu N., 2004. Genetic diversity of oil palm (*Elaeis guineensis* Jacq.) germplasm collections from Africa: implications for improvement and conservation of genetic resources. *Theor. Appl. Genet*, 108(7): 1274–1284.
- Hayes B. & Goddard M., 2010. Genome-wide association and genomic selection in animal breeding. *Genome*, 53(11): 876–883.
- Hayes B.J., Bowman P.J., Chamberlain A.J. & Goddard M.E., 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci*, 92(2): 433–443.
- 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. *Frontiers in plant science*, 5: 484–492.
- Heffner E.L., Sorrells M.E. & Jannink J.-L., 2009. Genomic selection for crop improvement. *Crop Science*, 49(1): 1–12.
- Herrero J., Santika B., Herrán A., Erika P., Sarimana U., Wendra F., Sembiring Z., Asmono D. & Ritter E., 2020. Construction of a high density linkage map in Oil Palm using SPET markers. *Scientific Reports*, 10(1): 1–9.

- Heslot N., Jannink J.-L. & Sorrells M., 2015. Perspectives for Genomic Selection Applications and Research in Plants. *Crop Science*, 55(1): 1–12.
- Hormaza P., Fuquen E.M. & Romero H.M., 2012. Phenology of the oil palm interspecific hybrid *Elaeis oleifera* × *Elaeis guineensis*. *Sci. agric.(Piracicaba, Braz.)*. 69: 275–280.
- Hufford M.B., Lubinsky P., Pyhäjärvi T., Devengenzo M.T., Ellstrand N.C. & Ross-Ibarra J., 2013. The Genomic Signature of Crop-Wild Introgression in Maize. *PLOS Genetics*, 9(5): e1003477.
- I M., Rajanaidu N., Zakri A.H. & Cheah S.C., 2006. Assessment of Genetic Diversity in Oil Palm (*Elaeis guineensis* Jacq.) using Restriction Fragment Length Polymorphism (RFLP). *Genet Resour Crop Evol*, 53:187–195.
- Imai A., Kuniga T., Yoshioka T., Nonaka K., Mitani N., Fukamachi H., Hiehata N., Yamamoto M. & Hayashi T., 2019. Single-step genomic prediction of fruit-quality traits using phenotypic records of non-genotyped relatives in citrus. *PLOS ONE*, 14(8): e0221880.
- Intara Y.I., Nusantara A.D., Supanjani S., Caniago Z. & Ekawita R., 2018. Oil palm roots architecture in response to soil humidity. *International journal of oil palm*, 1: 79–89.
- Isidro J., Jannink J.-L., Akdemir D., Poland J., Heslot N. & Sorrells M.E., 2015. Training set optimization under population structure in genomic selection. *Theoretical and applied genetics*, 128(1): 145–158.
- Isidro y Sánchez J. & Akdemir D., 2021. Training set optimization for sparse phenotyping in genomic selection: A conceptual overview. *Frontiers in Plant Science*, 1889 p.
- Isik F., 2014. Genomic selection in forest tree breeding: the concept and an outlook to the future. *New Forests*, 45(3): 379–401.
- Ismail S.R., Maarof S.K., Siedar Ali S. & Ali A., 2018. Systematic review of palm oil consumption and the risk of cardiovascular disease. *PLoS One*, 13(2): e0193533.
- Ithnin M. & Din A.K., 2020. The Oil Palm Genome, Compendium of Plant Genomes. Springer International Publishing. 505 p.
- Ithnin M., The C.-K. & Ratnam W., 2017. Genetic diversity of *Elaeis oleifera* (HBK) Cortes populations using cross species SSRs: implication's for germplasm utilization and conservation. *BMC genetics*, 18(1): 1–12.
- Jacquemard J.-C., Baudouin L. & Noiret J.-M., 2001. Oil palm. *Tropical plant breeding* 338 p.
- Jakobsson M., Edge M.D. & Rosenberg N.A., 2013. The Relationship Between FST and the Frequency of the Most Frequent Allele. *Genetics*, 193(2): 515–528.
- Jalani B.S., Cheah S.C., Rajanaidu N. & Darus A., 1997. Improvement of palm oil through breeding and biotechnology. *J Amer Oil Chem Soc*, 74(11): 1451–1455.

- Jannink J.-L., Lorenz A.J. & Iwata H., 2010. Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics*, 9(2): 166–177.
- Jeennor S. & Volkaert H., 2014. Mapping of quantitative trait loci (QTLs) for oil yield using SSRs and gene-based markers in African oil palm (*Elaeis guineensis* Jacq.). *Tree Genetics & Genomes*, 10(1): 1–14.
- Jemaa S.B., Thamri N., Mnara S., Rebours E., Rocha D. & Boussaha M., 2019. Linkage disequilibrium and past effective population size in native Tunisian cattle. *Genet. Mol. Biol*, 42(1): 52–61.
- Jensen S.M., Svensgaard J. & Ritz C., 2020. Estimation of the harvest index and the relative water content—Two examples of composite variables in agronomy. *European Journal of Agronomy*, 112: 125962.
- Jin J., Lee M., Bai B., Sun Y., Qu J., Rahmadsyah null, Alfiko Y., Lim C.H., Suwanto A., Sugiharti M., Wong L., Ye J., Chua N.-H. & Yue G.H., 2016. Draft genome sequence of an elite Dura palm and whole-genome patterns of DNA variation in oil palm. *DNA Res*, 23(6): 527–533.
- John Martin J.J., Yarra R., Wei L. & Cao H., 2022. Oil Palm Breeding in the Modern Era: Challenges and Opportunities. *Plants*, 11: 1395.
- Johnston H.R., Keats B.J.B. & Sherman S.L., 2019. 12 - Population Genetics, in: Pyeritz, R.E., Korf, B.R., Grody, W.W. (Eds.), Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics (Seventh Edition). Academic Press, 359–373 p.
- Jourdan C. & Rey H., 1997. Modelling and simulation of the architecture and development of the oil-palm (*Elaeis guineensis* Jacq.) root system. *Plant and Soil*, 190: 235–246.
- Kadandale S., Marten R. & Smith R., 2019. The palm oil industry and noncommunicable diseases. *Bulletin of the World Health Organization*. 97(2):118-128.
- Kijas J.W., Lenstra J.A., Hayes B., Boitard S., Neto L.R.P., Cristobal M.S., Servin B., McCulloch R., Whan V., Gietzen K., Paiva S., Barendse W., Ciani E., Raadsma H., McEwan J., Dalrymple B., Consortium. & other members of the I.S.G., 2012. Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *PLOS Biology*, 10(2): e1001258.
- Kimura M., 1983. The neutral theory of molecular evolution. Cambridge University Press.
- Kliman R., Sheehy B. & Schultz J., 2008. Genetic drift and effective population size. *Nature Education*, 1(3): 3.
- Kumar B., Rakshit S., Kumar S., Singh B.K., Lahkar C., Jha A.K., Kumar K., Kumar P., Choudhary M. & Singh S.B., 2022. Genetic Diversity, Population Structure and Linkage

- Disequilibrium Analyses in Tropical Maize Using Genotyping by Sequencing. *Plants*, 11(6): 799.
- Kwong Q.B., Ong A.L., Teh C.K., Chew F.T., Tammi M., Mayes S., Kulaveerasingham 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.). *Sci Rep*, 7(1): 1–9.
- Kwong Q.B., Teh C.K., Ong A.L., Chew F.T., Mayes S., Kulaveerasingham 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 Genetics*, 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., Kulaveerasingham H., Tammi M. & Appleton D.R., 2016. Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm. *Molecular Plant*, 9(8): 1132–1141.
- Langmead B. & Salzberg S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4): 357–359.
- Lebedev V.G., Lebedeva T.N., Chernodubov A.I. & Shestibratov K.A., 2020. Genomic Selection for Forest Tree Improvement: Methods, Achievements and Perspectives. *Forests*, 11(11): 1190.
- Lee M., Xia J.H., Zou Z., Ye J., Rahmadsyah Alfiko Y., Jin J., Lieando J.V., Purnamasari M.I., Lim C.H., Suwanto A., Wong L., Chua N.-H. & Yue G.H., 2015. A consensus linkage map of oil palm and a major QTL for stem height. *Scientific Reports*, 5(1): 1–7.
- Lenz P.R.N., Beaulieu J., Mansfield S.D., Clément S., Desponts M. & Bousquet J., 2017. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (*Picea mariana*). *BMC Genomics*, 18(1): 1–17.
- Lewontin R.C. & Kojima K., 1960. The Evolutionary Dynamics of Complex Polymorphisms. *Evolution*, 14: 458–472.
- Li Y. & Kim J.-J., 2015. Effective population size and signatures of selection using bovine 50K SNP chips in Korean native cattle (Hanwoo). *Evolutionary Bioinformatics*, 11: 143–53.
- Lin Z., Hayes B. & Daetwyler H., 2014. Genomic selection in crops, trees and forages: a review. *Crop and Pasture Science*, 65(11): 1177–1191.

- Liu X., Wang Hongwu., Wang Hui., Guo Z., Xu X., Liu J., Wang S., Li W.-X., Zou C., Prasanna B.M., Olsen M.S., Huang C. & Xu Y., 2018. Factors affecting genomic selection revealed by empirical evidence in maize. *The Crop Journal*, 6(4): 341–352.
- Lloyd S.S., Steele E.J. & Dawkins R.L., 2016. Analysis of Haplotype Sequences. Next Generation Sequencing - Advances, Applications and Challenges. *InTechOpen*. 345–368 p.
- Lourenco D., Legarra A., Tsuruta S., Masuda Y., Aguilar I. & Misztal I., 2020. Single-step genomic evaluations from theory to practice: using SNP chips and sequence data in BLUPF90. *Genes*, 11(7): 790.
- Mackay I. & Powell W., 2007. Methods for linkage disequilibrium mapping in crops. *Trends in plant science*, 12(2): 57–63.
- Maia R.T. & de Araújo Campos M., 2019. Introductory Chapter: Population Genetics-The Evolution Process as a Genetic Function, *In: Integrated View of Population Genetics*. IntechOpen. 5 p.
- Makina S.O., Taylor J.F., van Marle-Köster E., Muchadeyi F.C., Makgahlela M.L., MacNeil M.D. & Maiwashe A., 2015. Extent of Linkage Disequilibrium and Effective Population Size in Four South African Sanga Cattle Breeds. *Front. Genet.* 6: 337.
- Maldonado C., Mora F., Contreras-Soto R., Ahmar S., Chen J.-T., do Amaral Júnior A.T. & Scapim C.A., 2020. Genome-wide prediction of complex traits in two outcrossing plant species through Deep Learning and Bayesian Regularized Neural Network. *Frontiers in Plant Science*, 11: 1734.
- Maldonado C., Mora F., Scapim C.A. & Coan M., 2019. Genome-wide haplotype-based association analysis of key traits of plant lodging and architecture of maize identifies major determinants for leaf angle: Hap LA4. *PloS one*, 14(3): e0212925.
- Mancini A., Imperlini E., Nigro E., Montagnese C., Daniele A., Orrù S. & Buono P., 2015. Biological and nutritional properties of palm oil and palmitic acid: effects on health. *Molecules*, 20(9): 17339–17361.
- Marchal A., Legarra A., Tisne S., Carasco-Lacombe C., Manez A., Suryana E., Omoré A., Nouy B., Durand-Gasselin T. & Sánchez L., 2016. Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests. *Molecular Breeding*, 36(1): 1–13.
- Mayes S., Jack P.L., Corley R.H. & Marshall D.F., 1997. Construction of a RFLP genetic linkage map for oil palm (*Elaeis guineensis* Jacq.). *Genome*, 40(1): 116–122.

- Mba O.I., Dumont M.-J. & Ngadi M., 2015. Palm oil: Processing, characterization and utilization in the food industry—A review. *Food bioscience*, 10: 26–41.
- McElroy M.S., Navarro A.J.R., Mustiga G., Stack C., Gezan S., Peña G., Sarabia W., Saquicela D., Sotomayor I., Douglas G.M., Migicovsky Z., Amores F., Tarqui O., Myles S. & Motamayor J.C., 2018. Prediction of Cacao (*Theobroma cacao*) Resistance to *Moniliophthora* spp. Diseases via Genome-Wide Association Analysis and Genomic Selection. *Front. Plant Sci*, 9: 343.
- Meijaard E., Brooks T.M., Carlson K.M., Slade E.M., Garcia-Ulloa J., Gaveau D.L., Lee J.S.H., Santika T., Juffe-Bignoli D. & Struebig M.J., 2020. The environmental impacts of palm oil in context. *Nature plants*, 6:1418–1426.
- Mergeai G., Baudoin J. P., Demol J., Louant B. P., Maréchal R. & Otoul É., 2002. Plant breeding: its application to the main species grown in tropical regions. Presses Agronomiques de Gembloux. 582 p.
- Meunier J., 1969. Etude des populations naturelles d'*Elaeis guineensis* en Côte-d'Ivoire. *Oléagineux*, 24(4): 195–201.
- Meunier J. & Hardon J., 1976. Interspecific hybrids between *Elaeis guineensis* and *Elaeis oleifera*. *Oil palm research*, 14(1): 127–138.
- 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.
- Minamikawa M.F., Nonaka K., Kaminuma E., Kajiya-Kanegae H., Onogi A., Goto S., Yoshioka T., Imai A., Hamada H. & Hayashi T., 2017. Genome-wide association study and genomic prediction in citrus: potential of genomics-assisted breeding for fruit quality traits. *Scientific reports*, 7(1): 1–13.
- Momen M., Mehrgardi A.A., Sheikhi A., Kranis A., Tusell L., Morota G., Rosa G.J.M. & Gianola D., 2018. Predictive ability of genome-assisted statistical models under various forms of gene action. *Scientific Reports*, 8(1): 1–11.
- Montesinos-López O.A., Montesinos-López A., Pérez-Rodríguez P., Barrón-López J.A., Martini J.W., Fajardo-Flores S.B., Gaytan-Lugo L.S., Santana-Mancilla P.C. & Crossa J., 2021. A review of deep learning applications for genomic selection. *BMC genomics*, 22(1): 1–23.
- Montoya C., Lopes R., Flori A., Cros D., Cuellar T., Summo M., Espeout S., Rivallan R., Risterucci A.-M., Bittencourt D., Zambrano J.R., Alarcón G W.H., Villeneuve P., Pina M., Nouy B., Amblard P., Ritter E., Leroy T. & Billotte N., 2013. Quantitative trait loci (QTLs) analysis of palm oil fatty acid composition in an interspecific pseudo-backcross

- from *Elaeis oleifera* (H.B.K.) Cortés and oil palm (*Elaeis guineensis* Jacq.). *Tree Genetics & Genomes*, 9(5): 1207–1225.
- Moretzsohn M.C., Nunes C.D.M., Ferreira M.E. & Grattapaglia D., 2000. RAPD linkage mapping of the shell thickness locus in oil palm (*Elaeis guineensis* Jacq.). *Theor Appl Genet*, 100: 63–70.
- Morota G. & Gianola D., 2014. Kernel-based whole-genome prediction of complex traits: a review. *Frontiers in genetics*, 5: 363.
- Mozzon M., Pacetti D., Lucci P., Balzano M. & Frega N.G., 2013. Crude palm oil from interspecific hybrid *Elaeis oleifera* × *Elaeis guineensis*: Fatty acid regiodistribution and molecular species of glycerides. *Food Chemistry*, 141(1): 245–252.
- Mphahlele M.M., Isik F., Hodge G.R. & Myburg A.A., 2021. Genomic breeding for diameter growth and tolerance to *Leptocybe* gall wasp and *Botryosphaeria*/*Teratosphaeria* fungal disease complex in *Eucalyptus grandis*. *Frontiers in plant science*, 12: 228.
- Müller B.S.F., Neves L.G., de Almeida Filho J.E., Resende M.F.R., Muñoz P.R., dos Santos P.E.T., Filho E.P., Kirst M. & Grattapaglia D., 2017. Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of *Eucalyptus*. *BMC Genomics*, 18(1): 1–17.
- Munyengwa N., Le Guen V., Bille H.N., Souza L.M., Clément-Demange A., Mournet P., Masson A., Soumahoro M., Kouassi D. & Cros D., 2021. Optimizing imputation of marker data from genotyping-by-sequencing (GBS) for genomic selection in non-model species: Rubber tree (*Hevea brasiliensis*) as a case study. *Genomics*, 113(2): 655–668.
- 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 Genetics & Genomes*, 10(6): 1491–1510.
- Murphy D., 2014. The future of oil palm as a major global crop: Opportunities and challenges. *Journal of Oil Palm Research*, 26: 1–24.
- Murphy D., Goggin K. & Paterson R., 2020. Oil Palm in the 2020s and Beyond: Challenges and Solutions. *CABI Agriculture and Bioscience*, 2(1):1–22.
- Nagylaki T., 1998. Fixation Indices in Subdivided Populations. *Genetics*, 148(3): 1325–1332.
- Nainggolan M. & Sinaga A.G.S., 2021. Characteristics of fatty acid composition and minor constituents of red palm olein and palm kernel oil combination. *Journal of Advanced Pharmaceutical Technology & Research*, 12: 22.
- Nair K.P., 2021. Oil palm (*Elaeis guineensis* Jacquin), In: *Tree Crops*. Springer, 249–285 p.

- Nair K.P., 2010. Oil palm (*Elaeis guineensis* Jacquin). In: The Agronomy and Economy of Important Tree Crops of the Developing World. ELSEVIER INSIGHTS. 209-235 p.
- Nakaya A. & Isobe S.N., 2012. Will genomic selection be a practical method for plant breeding? *Ann Bot*, 110(6): 1303–1316.
- Namkoong G., Kang H.C. & Brouard J.S., 2012. Tree Breeding: Principles and Strategies: Principles and Strategies. Springer Science & Business Media. 11 p.
- Ngalle H.B., 2016. Nouveau schéma de valorisation de la mutation au locus sh pour la productivité de l'huile de palme chez *Elaeis guineensis* Jacq (Thèse de Doctorat). Université de Yaoundé I, Cameroon. 18 p.
- Nielsen N.H., Jahoor A., Jensen J.D., Orabi J., Cericola F., Edriss V. & Jensen J., 2016. Genomic prediction of seed quality traits using advanced barley breeding lines. *PLoS One*, 11(10): e0164494.
- Nyine M., Uwimana B., Blavet N., Hřibová E., Vanrespaille H., Batte M., Akech V., Brown A., Lorenzen J., Swennen R. & Doležel J., 2018. Genomic Prediction in a Multiploid Crop: Genotype by Environment Interaction and Allele Dosage Effects on Predictive Ability in Banana. *The Plant Genome*, 11(2): 170090.
- Nyouma A., 2021. Contribution of genomic selection to genetic improvement of palm oil yield in oil palm (*Elaeis guineensis* Jacq.) (PhD Thesis, Plant Biotechnology). University of Yaounde I, Cameroon. 53-54p.
- 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 Genetics & Genomes*, 15(5): 1–16.
- Nyouma A., Bell J.M., Jacob F., Riou V., Manez A., Pomiès V., Domonhede H., Arifiyanto D., Cochard B., Durand-Gasselín T. & Cros D., 2022. Improving the accuracy of genomic predictions in an outcrossing species with hybrid cultivars between heterozygote parents: a case study of oil palm (*Elaeis guineensis* Jacq.). *Molecular Genetics and Genomics*, 297(2): 523–533.
- Nyouma A., Bell J.M., Jacob F., Riou V., Manez A., Pomiès V., Nodichao L., Syahputra I., Affandi D., Cochard B., Durand-Gasselín T. & Cros, D., 2020. Genomic predictions improve clonal selection in oil palm (*Elaeis guineensis* Jacq.) hybrids. *Plant Science*, 299: 110547.
- Ojeda M., Borrero M., Sequeda G., Diez O., Castro V., García Á., Ruiz Á., Pacetti D., Frega N., Gagliardi R. & Lucci P., 2017. Hybrid palm oil (*Elaeis oleifera* × *Elaeis guineensis*)

- supplementation improves plasma antioxidant capacity in humans. *European Journal of Lipid Science and Technology*, 119(2): 1600070.
- Ong A.-L., Teh C.-K., Kwong Q.-B., Tangaya P., Appleton D.R., Massawe F. & Mayes S., 2019. Linkage-based genome assembly improvement of oil palm (*Elaeis guineensis*). *Scientific reports*, 9(1): 1–9.
- Ong A.-L., Teh C.-K., Mayes S., Massawe F., Appleton D.R. & Kulaveerasingam H., 2020. An improved oil palm genome assembly as a valuable resource for crop improvement and comparative genomics in the Arecoideae subfamily. *Plants*, 9(11): 1476.
- Ouellette L.A., Reid R.W., Blanchard S.G. & Brouwer C.R., 2018. LinkageMapView—rendering high-resolution linkage and QTL maps. *Bioinformatics*, 34(2): 306–307.
- Paludeto J.G.Z., Grattapaglia D., Estopa R.A. & Tambarussi E.V., 2021. Genomic relationship-based genetic parameters and prospects of genomic selection for growth and wood quality traits in *Eucalyptus benthamii*. *Tree Genetics & Genomes*, 17(4): 1–20.
- Pan B., Kusko R., Xiao W., Zheng Y., Liu Z., Xiao C., Sakkiah S., Guo W., Gong P. & Zhang C., 2019. Similarities and differences between variants called with human reference genome HG19 or HG38. *BMC bioinformatics*, 20(2): 17–29.
- Paterson R.R.M. & Lima N., 2018. Climate change affecting oil palm agronomy, and oil palm cultivation increasing climate change, require amelioration. *Ecology and Evolution*, 8(1): 452–461.
- Peixoto L. de A., Laviola B.G., Alves A.A., Rosado T.B. & Bhering L.L., 2017. Breeding *Jatropha curcas* by genomic selection: A pilot assessment of the accuracy of predictive models. *PLOS ONE*, 12(3): e0173368.
- Pirker J., Mosnier A., Kraxner F., Havlík P. & Obersteiner M., 2016. What are the limits to oil palm expansion? *Global Environmental Change*, 40: 73–81.
- Plavšín I., Gunjača J., Šatović Z., Šarčević H., Ivić M., Dvojković K. & Novoselović D., 2021. An Overview of Key Factors Affecting Genomic Selection for Wheat Quality Traits. *Plants*, 10(4): 745.
- Poets A.M., Fang Z., Clegg M.T. & Morrell P.L., 2015. Barley landraces are characterized by geographically heterogeneous genomic origins. *Genome Biology*, 16(1): 173.
- Pootakham W., Jomchai N., Ruang-areerate P., Shearman J.R., Sonthirod C., Sangsrakru D., Tragoonrung S. & Tangphatsornruang S., 2015. Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics*, 105(5-6): 288–295.

- Porto-Neto L.R., Lee S.H., Lee H.K. & Gondro C., 2013. Detection of signatures of selection using F_{ST}, in: *Genome-Wide Association Studies and Genomic Prediction*. Springer, 423–436 p.
- 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. *Journal of dairy science*, 95(1): 389–400.
- Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., De Bakker P.I. & Daly M.J., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3): 559–575.
- Purseglove J.W., 1986. *Tropical Crops -Monocotyledons*. Wiley, Harlow, Essex, England : New York. 291–310 p.
- Qaim M., Sibhatu K.T., Siregar H. & Grass I., 2020. Environmental, economic, and social consequences of the oil palm boom. *Annual review of resource economics*, 12: 321–344.
- Qian L., Hickey L.T., Stahl A., Werner C.R., Hayes B., Snowdon R.J. & Voss-Fels K.P., 2017. Exploring and Harnessing Haplotype Diversity to Improve Yield Stability in Crops. *Front Plant Sci*, 8: 1534.
- R Development Core Team., 2022. a language and environment for statistical computing: reference index. R Foundation for Statistical Computing, Vienna.
- Rahman M., Hoque A. & Roy J., 2022. Linkage disequilibrium and population structure in a core collection of *Brassica napus* (L.). *PloS one*, 17(3): e0250310.
- Rajinder S. & Choo C., 2005. Potential application of marker-assisted selection (MAS) in oil palm. *Oil Palm Bulletin*, 51: 1–9.
- Rambolarimanana T., Ramamonjisoa L., Verhaegen D., Tsy J.-M.L.P., Jacquin L., Cao-Hamadou T.-V., Makouanzi G. & Bouvet J.-M., 2018. Performance of multi-trait genomic selection for *Eucalyptus robusta* breeding program. *Tree Genetics & Genomes*, 14(5): 1–13.
- Rance K., Mayes S., Price Z., Jack P. & Corley R., 2001. Quantitative trait loci for yield components in oil palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics*, 103(8): 1302–1310.
- Rastas P., 2017. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics*, 33(23): 3726–3732.

- Reich D.E., Cargill M., Bolk S., Ireland J., Sabeti P.C., Richter D.J., Lavery T., Kouyoumjian R., Farhadian S.F., Ward R. & Lander E.S., 2001. Linkage disequilibrium in the human genome. *Nature*, 411(6834): 199–204.
- Resende M.D.V., Resende M.F.R., Sansaloni C.P., Petroli C.D., Missiaggia A.A., Aguiar A.M., Abad J.M., Takahashi E.K., Rosado A.M., Faria D.A., Pappas G.J., Kilian A. & Grattapaglia D., 2012. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol*, 194: 116–128.
- Resende R.T., Resende M.D.V., Silva F.F., Azevedo C.F., Takahashi E.K., Silva-Junior O.B. & Grattapaglia D., 2017. Assessing the expected response to genomic selection of individuals and families in Eucalyptus breeding with an additive-dominant model. *Heredity (Edinb)*, 119(4): 245–255.
- Rival A., 2017. Breeding the oil palm (*Elaeis guineensis* Jacq.) for climate change. OCL. 24: D107.
- Rivas M., Barbieri R.L. & Maia L.C. da., 2012. Plant breeding and in situ utilization of palm trees. *Ciência Rural*. 42: 261–269.
- Robertsen C.D., Hjortshøj R.L. & Janss L.L., 2019. Genomic Selection in Cereal Breeding. *Agronomy*, 9(2): 95.
- Rodríguez J.C., Gómez D., Pacetti D., Núñez O., Gagliardi R., Frega N.G., Ojeda M.L., Loizzo M.R., Tundis R. & Lucci P., 2016. Effects of the Fruit Ripening Stage on Antioxidant Capacity, Total Phenolics, and Polyphenolic Composition of Crude Palm Oil from Interspecific Hybrid *Elaeis oleifera* × *Elaeis guineensis*. *J. Agric. Food Chem*, 64(4): 852–859.
- Romero Navarro J.A., Phillips-Mora W., Arciniegas-Leal A., Mata-Quirós A., Haiminen N., Mustiga G., Livingstone III D., van Bakel H., Kuhn D.N., Parida L., Kasarskis A. & Motamayor J.C., 2017. Application of Genome Wide Association and Genomic Prediction for Improvement of Cacao Productivity and Resistance to Black and Frosty Pod Diseases. *Front. Plant Sci*, 8: 1905 p.
- Röös E., Bajželj B., Smith P., Patel M., Little D. & Garnett T., 2017. Greedy or needy? Land use and climate impacts of food in 2050 under different livestock futures. *Global Environmental Change*, 47: 1–12.
- Rosas Urióstegui F.I., Pat Fernández J.M., Pat Fernández L.A., Cornelis van der Wal J., Rosas Urióstegui F.I., Pat Fernández J.M., Pat Fernández L.A. & Cornelis van der Wal J.,

2018. The effect of oil palm on income strategies and food security of households in rural communities in Campeche, Mexico. *Acta universitaria*, 28(2): 25–32.
- Sallam A.H., Conley E., Prakapenka D., Da Y. & Anderson J.A., 2020. Improving prediction accuracy using multi-allelic haplotype prediction and training population optimization in wheat. *G3: Genes, Genomes, Genetics*, 10(7): 2265–2273.
- Sbordoni V., Allegrucci G. & Cesaroni D., 2004. Population structure. 447–455 p.
- Semagn K., Bjørnstad Å. & Ndjiondjop M.N., 2006. Principles, requirements and prospects of genetic mapping in plants. *African Journal of Biotechnology*, 5(25): 2570–2585.
- Seng T.-Y., Ritter E., Mohamed Saad S.H., Leao L.-J., Harminder Singh R.S., Qamaruz Zaman F., Tan S.-G., Syed Alwee S.S.R. & Rao V., 2016. QTLs for oil yield components in an elite oil palm (*Elaeis guineensis*) cross. *Euphytica*, 212(3): 399–425.
- Seng T.-Y., Saad S.H.M., Chin C.-W., Ting N.-C., Singh R.S.H., Zaman F.Q., Tan S.-G. & Alwee S.S.R.S., 2011. Genetic Linkage Map of a High Yielding FELDA Deli×Yangambi Oil Palm Cross. *PLOS ONE*, 6 (11): e26593.
- Siberchicot A., Bessy A., Guéguen L. & Marais G.A., 2017. MareyMap Online: A User-Friendly Web Application and Database Service for Estimating Recombination Rates Using Physical and Genetic Maps. *Genome Biol Evol*, 9(10): 2506–2509.
- Silva F.A., Viana A.P., Corrêa C.C.G., Santos E.A., Oliveira J.A.V.S., Andrade J.D.G., Ribeiro R.M. & Glória L.S., 2021. Bayesian Ridge Regression Shows the Best Fit for Ssr Markers in Psidium Guajava Among Bayesian Models. *Scientific Reports*, 11(1): 1–11.
- Silva-Junior O.B., Faria D.A. & Grattapaglia D., 2015. A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. *New Phytologist*, 206(4): 1527–1540.
- 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., Nagappan J., Chan K.-L., Rosli R., Halim M.A., Azizi N., Budiman M.A., Lakey N., Bacher B., Van Brunt A., Wang C., Hogan M., He D., MacDonald J.D., Smith S.W., Ordway J.M., Martienssen R.A. & Sambanthamurthi R., 2014. The oil palm VIRESCENS gene controls fruit colour and encodes a R2R3-MYB. *Nature Communications*, 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.-E., Chan K.-L., Halim M.A., Azizi N., Nagappan J., Bacher B., Lakey N., Smith S.W., He D., Hogan M., Budiman M.A., Lee E.K., DeSalle R., Kudrna D., Goicoechea J.L., Wing R.A., Wilson R.K., Fulton R.S., Ordway J.M., Martienssen R.A. &

- Sambanthamurthi R., 2013. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature*, 500(7462): 335–339.
- Singh R., Tan S.G., Panandam J.M., Rahman R.A., Ooi L.C., Low E.-T.L., Sharma M., Jansen J. & Cheah S.-C., 2009. Mapping quantitative trait loci (QTLs) for fatty acid composition in an interspecific cross of oil palm. *BMC Plant Biol*, 9(1): 1–19.
- Slatkin M., 2008. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6): 477–485.
- Soh A.C., 2018. Applications and challenges of biotechnology in oil palm breeding. IOP Conf. Ser.: Earth Environ. Sci. 183(1): 012002.
- Soh A.C., 2012. 2 - Breeding and Genetics of the Oil Palm, in: Lai, O.-M., Tan, C.-P., Akoh, C.C. (Eds.), *Palm Oil*. AOCS Press, 31–58 p.
- Soh A.C., Gan H.H., Wong G., Hor T.Y. & Tan C.C., 2003. Estimates of within family genetic variability for clonal selection in oil palm. *Euphytica*, 133(2): 147–163.
- Soh A.C., Mayes S. & Roberts J.A. (Eds.), 2017. *Oil Palm Breeding: Genetics and Genomics*. CRC Press, Boca Raton, 446 p.
- Solberg T.R., Sonesson A.K., Woolliams J.A. & Meuwissen T.H.E., 2008. Genomic selection using different marker types and densities. *Journal of Animal Science*, 86(10): 2447–2454.
- Sorkheh K., Malysheva-Otto L.V., Wirthensohn M.G., Tarkesh-Esfahani S. & Martínez-Gómez P., 2008. Linkage disequilibrium, genetic association mapping and gene localization in crop plants. *Genet. Mol. Biol*, 31(4): 805–814.
- Sousa I.C. de, Nascimento M., Silva G.N., Nascimento A.C.C., Cruz C.D., Almeida D.P. de, Pestana K.N., Azevedo C.F., Zambolim L. & Caixeta E.T., 2020. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. *Scientia Agricola*, 78.
- Sousa T.V., Caixeta E.T., Alkimim E.R., Oliveira A.C.B., Pereira A.A., Sakiyama N.S., Zambolim L. & Resende M.D.V., 2019. Early Selection Enabled by the Implementation of Genomic Selection in Coffea arabica Breeding. *Front. Plant Sci*, 9:1934
- Souza L.M., Francisco F.R., Gonçalves P.S., Scaloppi Junior E.J., Le Guen V., Fritsche-Neto R. & Souza A.P., 2019. Genomic selection in rubber tree breeding: a comparison of models and methods for managing G× E interactions. *Frontiers in plant science*, 10: 1353.
- Squire G., 2005. *The Oil Palm*. Oxford: Blackwell Publishing, 284 p.

- Statista 2021. Vegetable oils: production worldwide 2012/13-2020/21, by type. <https://www.statista.com/statistics/263933/production-of-vegetable-oils-worldwide-since-2000/>, n.d. Accessed November, 2021.
- Steiger J.H., 1980. Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2): 245.
- Tan B., Grattapaglia D., Martins G.S., Ferreira K.Z., Sundberg B. & Ingvarsson P.K., 2017. Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC Plant Biology*, 17(1): 1–15.
- Tan B., Grattapaglia D., Wu H.X. & Ingvarsson P.K., 2018. Genomic relationships reveal significant dominance effects for growth in hybrid Eucalyptus. *Plant Science*, 267: 84–93.
- Tandon R., 2001. Pollination and Pollen-pistil Interaction in Oil Palm, *Elaeis guineensis*. *Annals of Botany*, 87(6): 831–838.
- Tapia J.F.D., Doliente S.S. & Samsatli S., 2021. How much land is available for sustainable palm oil? *Land Use Policy*, 102: 105187.
- Technow F., Riedelsheimer C., Schrag T.A. & Melchinger A.E., 2012. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theoretical and Applied Genetics*, 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.
- Teh C.-K., Ong A.-L., Kwong Q.-B., Apparow S., Chew F.-T., Mayes S., Mohamed M., Appleton D. & Kulaveerasingam H., 2016. Genome-wide association study identifies three key loci for high mesocarp oil content in perennial crop oil palm. *Scientific Reports*, 6(1):1–7.
- Teh H.F., Neoh B.K., Wong Y.C., Kwong Q.B., Ooi T.E.K., Ng T.L.M., Tiong S.H., Low J.Y.S., Danial A.D., Ersad M.A., Kulaveerasingam H. & Appleton D.R., 2014. Hormones, polyamines, and cell wall metabolism during oil palm fruit mesocarp development and ripening. *J Agric Food Chem*, 62(32): 8143–8152.
- Tenaillon M.I., Austerlitz F. & Tenaillon O., 2008. Apparent mutational hotspots and long distance linkage disequilibrium resulting from a bottleneck. *Journal of Evolutionary Biology*, 21(2): 541–550.

- Thistlethwaite F.R., Gamal El-Dien O., Ratcliffe B., Klápště J., Porth I., Chen C., Stoehr M.U., Ingvarsson P.K. & El-Kassaby Y.A., 2020. Linkage disequilibrium vs. pedigree: Genomic selection prediction accuracy in conifer species. *PLoS One*, 15(6): e0232201.
- Ting N.-C., Jansen J., Mayes S., Massawe F., Sambanthamurthi R., Ooi L.C.-L., Chin C.W., Arulandoo X., Seng T.-Y., Alwee S.S.R.S., Ithnin M. & Singh R., 2014. High density SNP and SSR-based genetic maps of two independent oil palm hybrids. *BMC Genomics*, 15(1): 1–11.
- Ting N.-C., Jansen J., Nagappan J., Ishak Z., Chin C.-W., Tan S.-G., Cheah S.-C. & Singh R., 2013. Identification of QTLs Associated with Callogenesis and Embryogenesis in Oil Palm Using Genetic Linkage Maps Improved with SSR Markers. *PLOS ONE*, 8(1): e53076.
- 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
- Tisné S., Pomiès V., Riou V., Syahputra I., Cochard B. & Denis M., 2017. Identification of Ganoderma Disease Resistance Loci Using Natural Field Infection of an Oil Palm Multiparental Population. *G3*, 7(6): 1683–1692.
- Tong H. & Nikoloski Z., 2021. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *Journal of plant physiology*, 257: 153354.
- Tonks A.J., Aplin P., Beriro D.J., Cooper H., Evers S., Vane C.H. & Sjögersten S., 2017. Impacts of conversion of tropical peat swamp forest to oil palm plantation on peat organic chemistry, physical properties and carbon stocks. *Geoderma*, 289: 36–45.
- Uitdewilligen J.G., Wolters A.-M.A., D'hoop B.B., Borm T.J., Visser R.G. & Van Eck H.J., 2013. A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PloS one*, 8(5): e62355.
- Ukoskit K., Chanroj V., Bhusudsawang G., Pipatchartlearnwong K., Tangphatsornruang S. & Tragoonrung S., 2014. Oil palm (*Elaeis guineensis* Jacq.) linkage map, and quantitative trait locus analysis for sex ratio and related traits. *Mol Breeding*, 33(2): 415–424.
- van der Werf J., 2013. Genomic selection in animal breeding programs. *Methods Mol. Biol*, 1019: 543–561.
- Varshney R.K., Graner A. & Sorrells M.E., 2005. Genomics-assisted breeding for crop improvement. *Trends Plant Sci*, 10(12): 621–630.

- Vossen H.A.M. van der, 1974. Towards more efficient selection for oil yield in the oil palm (*Elaeis guineensis* Jacquin). Wageningen University and Research. 823 p.
- Voss-Fels K.P., Cooper M. & Hayes B.J., 2019. Accelerating crop genetic gains with genomic selection. *Theor. Appl. Genet*, 132(3), 669–686.
- Wand M., 1995. KernSmooth: Functions for Kernel Smoothing Supporting, Wand & Jones, R package version 2: 23–20.
- Wang J., Santiago E. & Caballero A., 2016. Prediction and estimation of effective population size. *Heredity*, 117: 193–206.
- Wang X., Xu Y., Hu Z. & Xu C., 2018. Genomic selection methods for crop improvement: Current status and prospects. *The Crop Journal*, 6(4): 330–340.
- Waples R.S. & Do C., 2008. ldne: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, 8(4): 753–756.
- Watson K., Mayes S., Price Z., Jack P. & Corley R., 2001. Quantitative trait loci for yield components in oil palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics*, 103(8): 1302–1310.
- Weir B.S., 1979. Inferences about Linkage Disequilibrium. *Biometrics* 35: 235–254.
- White T.L., Adams W.T. & Neale D.B., 2007. Forest genetics. CABI, Wallingford, UK; Cambridge, MA. 149 p.
- Wich S.A., Gaveau D., Abram N., Ancrenaz M., Baccini A., Brend S., Curran L., Delgado R.A., Erman A. & Fredriksson G.M., 2012. Understanding the impacts of land-use policies on a threatened species: is there a future for the Bornean orang-utan? *PloS one*, 7(11): e49142.
- Wientjes Y.C.J., Veerkamp R.F. & Calus M.P.L., 2013. The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. *Genetics*, 193: 621–631.
- Woittiez L.S., van Wijk M.T., Slingerland M., van Noordwijk M. & Giller K.E., 2017. Yield gaps in oil palm: A quantitative review of contributing factors. *European Journal of Agronomy*, 83: 57–77.
- Wong C.K. & 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.
- Wright S., 1978. Evolution and the Genetics of Populations, Volume 4: Variability Within and Among Natural Populations. University of Chicago Press, Chicago. 4 p.
- Wright S., 1931. Evolution in Mendelian Populations. *Genetics*, 16: 97–159.

- Xia W., Luo T., Zhang W., Mason A.S., Huang D., Huang X., Tang W., Dou Y., Zhang C. & Xiao Y., 2019. Development of High-Density SNP Markers and Their Application in Evaluating Genetic Diversity and Population Structure in *Elaeis guineensis*. *Front. Plant Sci*, 10:130.
- Xu H. & Guan Y., 2014. Detecting Local Haplotype Sharing and Haplotype Association. *Genetics*, 197: 823–838.
- Xu Y., Wang X., Ding X., Zheng X., Yang Z., Xu C. & Hu Z., 2018. Genomic selection of agronomic traits in hybrid rice using an NCII population. *Rice*, 11(1): 1–10.
- Yadav S., Ross E.M., Aitken K.S., Hickey L.T., Powell O., Wei X., Voss-Fels K.P. & Hayes B.J., 2021. A linkage disequilibrium-based approach to position unmapped SNPs in crop species. *BMC genomics*, 22(1): 1–9.
- Yamamoto E., Matsunaga H., Onogi A., Kajiya-Kanegae H., Minamikawa M., Suzuki A., Shirasawa K., Hirakawa H., Nunome T. & Yamaguchi H., 2016. A simulation-based breeding design that uses whole-genome prediction in tomato. *Scientific reports*, 6(1): 1–11.
- Yan L., Hofmann N., Li S., Ferreira M.E., Song B., Jiang G., Ren S., Quigley C., Fickus E., Cregan P. & Song Q., 2017. Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC Genomics*, 18(1):1–11.
- Yates M., Bowles E. & Fraser D., 2019. Small population size and low genomic diversity have no effect on fitness in experimental translocations of a wild fish. *Proceedings of the Royal Society B*. 286: 20191989.
- Ye S., Song H., Ding X., Zhang Z. & Li J., 2020. Pre-selecting markers based on fixation index scores improved the power of genomic evaluations in a combined Yorkshire pig population. *animal*, 14(8): 1555–1564.
- Yue G.H., Ye B.Q. & Lee M., 2021. Molecular approaches for improving oil palm for oil. *Molecular Breeding*, 41: 1–17.
- Zeven A.C., 1964. On the Origin of the Oil Palm (*Elaeis Guineensis* Jacq.). *Grana Palynologica*, 5(1): 121–123.
- Zhang H., Yin L., Wang M., Yuan X. & Liu X., 2019. Factors Affecting the Accuracy of Genomic Selection for Agricultural Economic Traits in Maize, Cattle, and Pig Populations. *Front. Genet*, 10:189.

- Zhao J., Sauvage C., Bitton F. & Causse M., 2022. Multiple haplotype-based analyses provide genetic and evolutionary insights into tomato fruit weight and composition. *Horticulture research*, 9 p.
- Zhao Y., Mette M.F. & Reif J.C., 2015. Genomic selection in hybrid breeding. *Plant Breeding*, 134(1): 1–10.
- Zheng X., Levine D., Shen J., Gogarten S.M., Laurie C. & Weir B.S., 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24): 3326–3328.
- Zhong S., Dekkers J.C.M., Fernando R.L. & Jannink J.-L., 2009. Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. *Genetics*, 182(1): 355–364.
- Zhou L. & Holliday J.A., 2012. Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC genomics*, 13(1): 1–12.

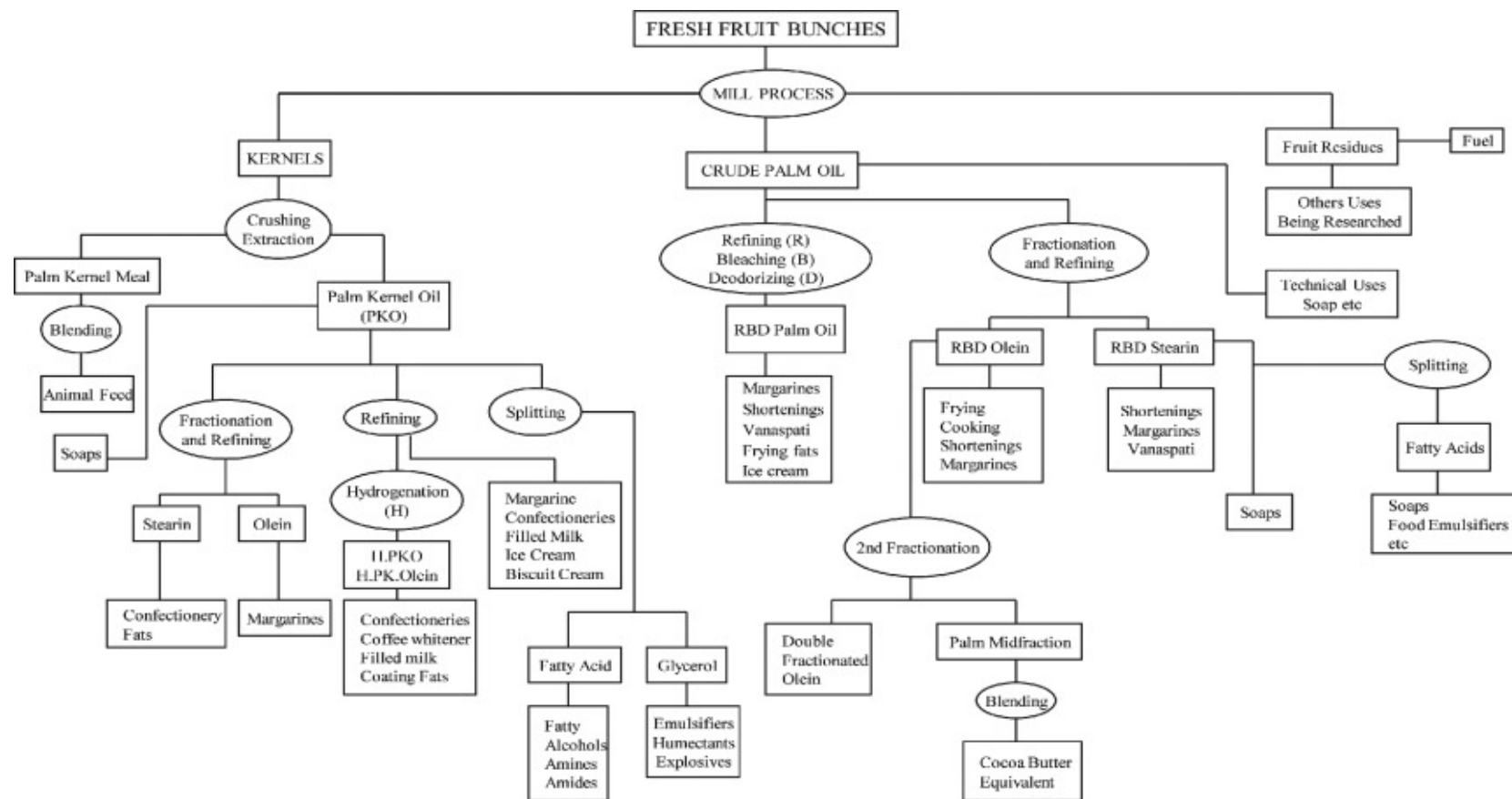
APPENDICES

APPENDICES

Appendix 1. Principal fatty acid compositions of the nine major globally traded vegetable oils (Murphy *et al.*, 2020).

Crop	% global supply	Principal fatty acids						
		12:0 Lauric	14:0 Myristic	16:0 Palmitic	18:0 Stearic	18:1 Oleic	18:2 Linoleic	18:3 α -Linolenic
Oil palm (mesocarp)	35.5		1	43	4	40	10	0.3
Oil palm (kernel)	4.3	48	16	8	2	15	25	
Soybean	27.8			11	4	23	54	8
Rapeseed	13.4			4	2	60	20	10
Sunflower	10.4			7	5	19	68	
Peanut	3.0			12	5	48	30	
Cottonseed	2.5		1	24	2	18	54	0.5
Coconut	1.8	49	17	9	2	6	2	
Olive	1.5			13	2	70	13	0.6

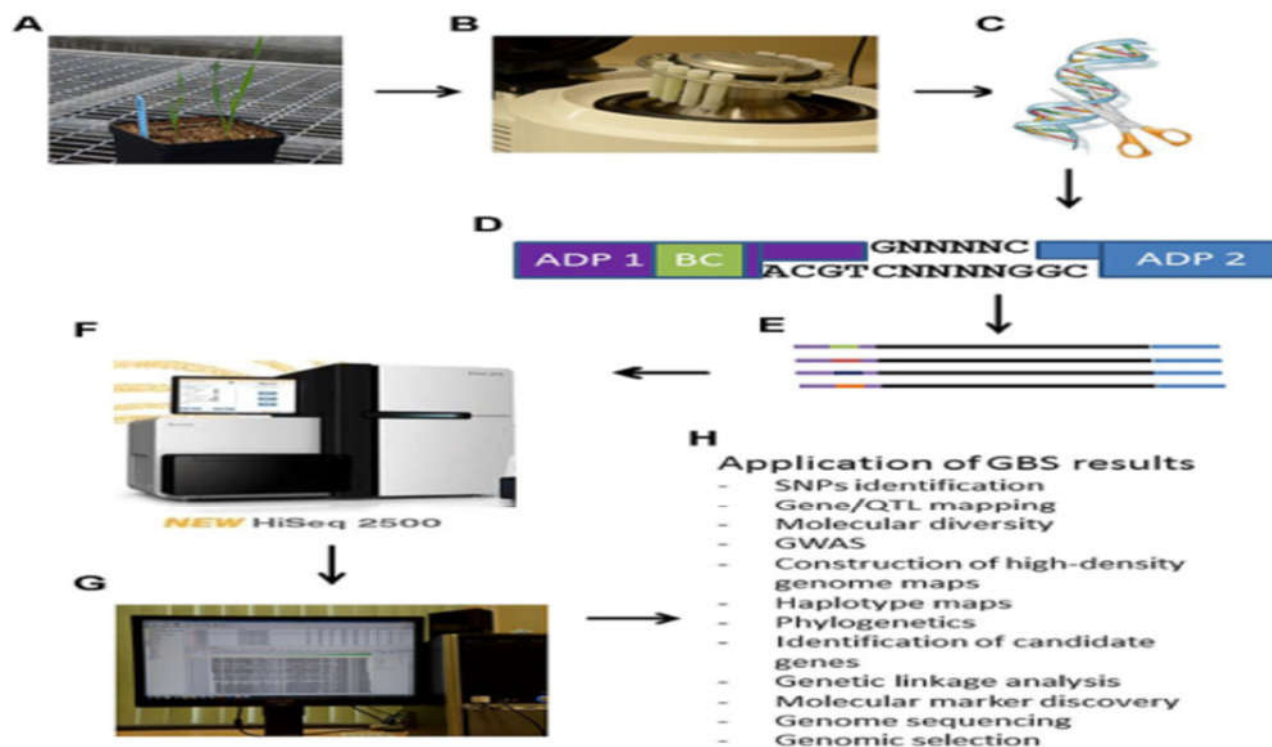
Appendix 2. Food and industrial importance of oil palm (Soh *et al.*, 2017).



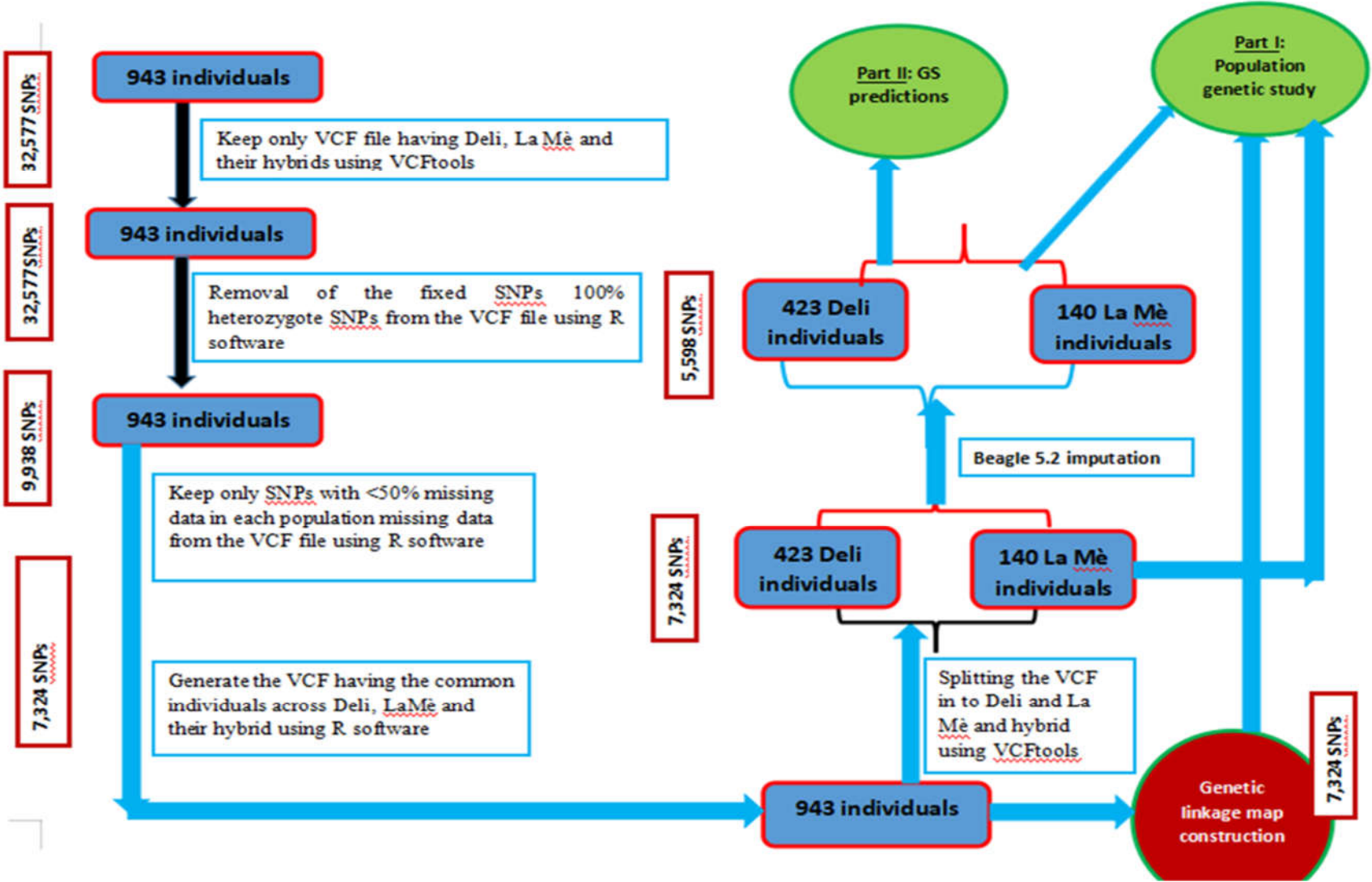
Appendix 3. Summary of linkage map constructed in oil palm.

No.	Year	Type of markers	No of Markers	Map depth (cM)	No of LGs	Software used	References
1	1997	RFLP	97	860	24	MAPMAKER 2.0	(Mayes <i>et al.</i> , 1997)
2	2000	RAPD	48	399.7-449.3	12-15	MAPMAKER 2.0	(Moretzsohn <i>et al.</i> , 2000)
3	2001	RFLP	153	852	22	JoinMap 2.0	(Rance <i>et al.</i> , 2001)
4	2005	MS and AFLP	255+688	1743	16	JoinMap ver. 3.0	(Billotte <i>et al.</i> , 2005)
5	2009	AFLP,RFLP,MS	252	1815	21	Joinmap ver. 4.0	(Singh <i>et al.</i> ,2009)
6	2010	SSR	251	1479	16	JoinMap v. 3.0	(Billotte <i>et al.</i> , 2010)
7	2011	AFLP	331	2274.5	16	MAPRF7	(Seng <i>et al.</i> , 2011)
8	2013	AFLP, RFLP, SSR	148	798.0	23	JoinMap 4.0	(Ting <i>et al.</i> , 2013)
9	2013	SSR	362	1845.0	16	JoinMap v.4.0	(Montoya <i>et al.</i> , 2013)
10	2014	SSR,Genes,SNP	190	1233.0	31	JoinMap v.4.0	(Jeennor&Volkaert, 2014)
11	2014	SSR,AFLP	423	1931	16	JoinMap v 3.0	(Ukoskit <i>et al.</i> ,2014)
12	2014	SSR, SNP	1331	1867	16	JoinMap v 4.1	(Ting <i>et al.</i> , 2014)
13	2015	SSR,SNP	480	1565.6	16	JoinMap v 3.0	(Lee <i>et al.</i> , 2015)
14	2015	SNP	1085	1429.6	16	JoinMap v 3.0	(Pootakham <i>et al.</i> , 2015)
15	2015	SSR	281	1935.0	16	CRIMAP	(Cochard <i>et al.</i> , 2015)
16	2018	SNP, SSR	10023	2398.2	16	Lep-MAP v 2	(Bai <i>et al.</i> , 2018a)
17	2018	SNP, DArT)based	1399-1466	1873.7-1720.6	16	JoinMap v 4.1	(Gan <i>et al.</i> , 2018)
18	2018	SNP	2413	1161.89	16	JoinMap v 4.1	(Bai <i>et al.</i> , 2018b)
19	2019	SNPs	27890	1151.7	16	Lep-MAP3	(Ong <i>et al.</i> ,2019)
20	2020	SPET	3,501	1,370	16	Lep-MAP3	(Herrero <i>et al.</i> , 2020)
21	2020	SNPs	11,421	1151.70- 1268.26	17- 24	Lep-MAP3	(Ong <i>et al.</i> , 2020)

Appendix 4. Major steps of genotyping-by-sequencing (GBS) protocol used in plant breeding. Step (a). Tissue is obtained from any plant (Elshire et al., 2011). Step (a). Tissue is obtained from any plant species; Step (b). Ground leaf tissues for DNA isolation, quantification and normalization. N.B: Take care of any cross-contamination among samples at this step; Step (c). DNA digestion with restriction enzymes; Step (d). Ligations of adaptors (e). including a bar coding (BC) region in adapter 1 in random PstI-MseI restricted DNA fragments; Step Representation of different amplified (f). DNA fragments with different bar codes from different biological samples/lines. N.B: These fragments represent the GSB library; (g). Analysis of sequences from library on a NGS sequencer; Step Bioinformatic analysis of NGS sequencing data; Step (H): Application of GBS results in breeding



Appendix 5. Generation of SNPs working.



Appendix 6. Objectives and corresponding published papers.

Objectives		Published papers
general	specific	
to characterize the genome properties of Deli and La Mé oil palm breeding populations for better palm oil yield.	to evaluate the genetic diversity between Deli and La Mé oil palm breeding populations	Genome properties of key oil palm (<i>Elaeis guineensis</i> Jacq.) breeding populations
	to estimate within-population linkage disequilibrium between Deli and La Mé oil palm breeding populations	
	to assess haplotype sharing between Deli and La Mé oil palm breeding populations	
	to determine the effective size between Deli and La Mé oil palm breeding populations	
	Genotyping by Sequencing for Plant Breeding- A Review	
	Genome Mapping to Enhance Efficient Marker-Assisted Selection and Breeding of the Oil Palm (<i>Elaeis guineensis</i> Jacq.)	
	Genomic selection in perennial tropical fruit and tree crops: a review	

Appendix 7. The logical framework of the specific objective I: evaluation of the genetic diversity between Deli and La Mé oil palm breeding populations.

Materials	Methods	Results	Conclusion
<p>The plant material used in this experiment consisted of individuals of the Deli, La Mé, and their crosses, i.e, 423 Deli, 140 La Mé, and 380 Deli × La Mé.</p> <p>We used a total of 7,324 SNP markers common for both breeding populations.</p>	<p>Molecular data were obtained by GBS</p> <p>Genome alignment: Bowtie2 software</p> <p>Sequence data were processed Tassel GBS and Vcftools</p> <p>We kept only Biallelic SNPs variants SNPs with 100% heterozygote genotypes were discarded Individuals with more than 50% missing data were removed Minor allele frequency (MAF), and heterozygosity between Deli and La Mé were obtained using 7,324 SNPs in R software</p> <p>Pairwise Fst between Deli and La Mé was estimated according to the Wright method using 7,324 SNPs and subsets of 100 random individuals per population using the SNPRelate R package</p>	<p>The average MAF was 0.09 for Deli and 0.14 for La Mé. Thus, the percentage of SNPs with MAF <0.05 was 60.5% in Deli and 49.7% in La Mé. The percentage of heterozygosity per individual ranged from 1.9% (Deli) to 20.9% (La Mé) The F_{st} between Deli and La Mé was 0.53. Several regions of the genome had high F_{st} values (>0.4), in particular on chromosomes EG51_2, EG51_8, and EG51_13 The correlation of heterozygosity per SNPs between the two populations showed that the majority of SNPs were, in one population, fixed or almost fixed The correlation in the frequency of alternate alleles per SNP between populations demonstrated that most SNPs were fixed or almost fixed with the reference allele in one population</p>	<p>With the high F_{st} value (0.53), the pattern of correlation of SNP heterozygosity and allele frequency among populations showed a significant degree of differentiation among the Deli and La Mé oil palm breeding populations.</p>

Appendix 8. The logical framework of the specific objective II: estimation within-population linkage disequilibrium for Deli and La Mé oil palm breeding populations.

Materials	Methods	Results	Conclusion
<p>The plant material used in this experiment consisted of individuals of the Deli, La Mé, and their crosses, i.e, 423 Deli, 140 La Mé, and 380 Deli × La Mé.</p> <p>For the construction of the genetic map, the 943 genotyped individuals of the breeding populations and their crosses were used, comprised on their pedigree, making a total of 1,788 individuals</p> <p>We used a total of 7,324 SNP markers common for both breeding populations, including 5,598 SNPs located on the assembled parts of the genome (i.e. the 16 chromosomes)</p>	<p>Molecular data were obtained by GBS Genome alignment: Bowtie2 software Sequence data were processed Tassel GBS and Vcftools We kept only Biallelic SNPs variants SNPs with 100% heterozygote genotypes were discarded. Individuals with more than 50% missing data were removed. Imputation of missing SNP data and phasing was carried out with Beagle 5.1. The genetic map was constructed using LepMAP3 software. The genetic position of the molecular markers against their physical position was plotted using MareyMap software.</p> <p>The linkage disequilibrium (LD) was performed in each breeding population using the PLINK software.</p> <p>The r^2 values between pairs of SNPs were plotted against physical distances (Mbp) and genetic distances (cM).</p> <p>The persistence of LD between populations was measured by the correlation of the r measure of LD between populations along with the genetic and physical maps using PLINK software.</p>	<p>The genetic map comprised 4,252 SNPs, spread over 2,782 unique positions, and spanned 1,778.52 cM with an average mapping interval between adjacent SNPs of 0.67 cM.</p> <p>The recombination rate was 2.85 cM/Mbp on average, ranging from 1.78 cM/Mbp (LG15) to 3.87 cM/Mbp.</p> <p>The LD reached high values (>0.6) for short distances between SNPs.</p> <p>It was higher in Deli than in La Mé for all distances.</p> <p>The corresponding distance between SNPs ($r^2=0.3$), was 1.05 cM in Deli and 0.9 cM in La Mé.</p> <p>The distance corresponding to SNPs ($r^2=0.3$) was 0.22 Mbp in Deli and 0.21 Mbp in La Mé.</p> <p>A high correlation of r values between populations (r_{LD}) was observed for close markers, i.e. r_{LD} above 0.6 for SNPs separated by a distance <0.5 cM on the genetic map or <1 kbp on the physical map.</p>	<p>The LD at $r^2=0.3$, considered as the minimum to get reliable results for genomic predictions, spanned over 1.05 cM/0.22 Mbp in Deli and 0.9 cM/0.21 Mbp in La Mé.</p> <p>$r^2=0.3$ we require around 1,700 SNPs for Deli and 2,000 SNPs for La Mé and when considering the physical distance we require around 2,900 SNPs in Deli and 83,100 SNPs in La Mé In the two populations, 10,000 SNPs would be enough to reach this level of LD with whole oil palm genome size.</p> <p>A high correlation of r values of LD between populations ($r_{LD}>0.6$) was obtained considering the markers separated by short distances, i.e., <0.5 cM on the genetic map or <1 kbp on the physical map.</p> <p>The level of resemblance between them over short genomic distances likely explains the superiority of GS models ignoring the parental origin of marker alleles</p> <p>Generally, the finding indicated that there is a strong genetic differentiation between Deli and La Mé</p>

Appendix 9. The logical framework of the specific objective III: assessment of the haplotype sharing between Deli and La Mé oil palm breeding populations.

Materials	Methods	Results	Conclusion
<p>The plant material used in this experiment consisted of individuals of the Deli, La Mé, and their crosses, i.e, 423 Deli, 140 La Mé, and 380 Deli × La Mé.</p> <p>For the construction of the genetic map, the 943 genotyped individuals of the breeding populations and their crosses were used, comprised on their pedigree, making a total of 1,788 individuals.</p> <p>We used a total of 7,324 SNP markers common for both breeding populations, including 5,598 SNPs located on the assembled parts of the genome (i.e. the 16 chromosomes).</p>	<p>Molecular data were obtained by GBS.</p> <p>Genome alignment: Bowtie2 software.</p> <p>Sequence data were processed Tassel GBS and Vcftools.</p> <p>We kept only Biallelic SNPs variants</p> <p>SNPs with 100% heterozygote genotypes were discarded</p> <p>Individuals with more than 50% missing data were removed</p> <p>Imputation of missing SNP data and phasing was carried out with Beagle 5.1.</p> <p>This analysis was done with the phased SNP data.</p> <p>Fifteen window sizes were used for physical distances, from 10 Mbp to 100 bp, and seven window sizes were used for genetic distances, from 10 cM to 0.01 cM.</p> <p>This analysis was done using a custom R script.</p>	<p>A large proportion of haplotypes were common between pairs of populations when considering short distances</p> <p>50% of the haplotypes with a length around 30 bp</p> <p>40% of the haplotypes with lengths around 3,600 bp were common to the two populations</p> <p>Fast haplotypes fall below 20% for haplotypes longer than 300 kbp.</p> <p>40% of the haplotypes with lengths around 0.20 cM were common to the two populations</p> <p>The length of the haplotypes increased, and the percentage of shared haplotypes between populations decreased.</p> <p>Fast haplotypes fall below 20% for haplotypes longer than 2.5 cM.</p> <p>The frequency of the common haplotype is less in short distances and high in long distances.</p>	<p>The percentage of common haplotypes was above 40% for short haplotypes (3600 bp or 0.20 cM). This resemblance decreased with the distance between SNPs, with for example the percentage of common haplotypes falling below 20% for haplotypes longer than 300 kbp.</p> <p>The level of resemblance between the two populations over short genomic distances (i.e, percentage of common haplotypes >40% for haplotypes <3600 bp/0.20 cM) likely explains the superiority of GS models ignoring the parental origin of marker alleles over models taking this information into account.</p> <p>The haplotype sharing with increasing SNP distance showed a significant degree of differentiation among the Deli and La Mé oil palm breeding populations.</p>

Appendix 10. The logical framework of the specific objective IV: evaluation of effective population size between Deli and La Mé.

Materials	Methods	Results	Conclusion
<p>The plant material used in this experiment consisted of individuals of the Deli, La Mé, and their crosses, i.e, 423 Deli, 140 La Mé, and 380 Deli × La Mé.</p> <p>For the construction of the genetic map, the 943 genotyped individuals of the breeding populations and their crosses were used, comprised on their pedigree, making a total of 1,788 individuals.</p> <p>We used a total of 7,324 SNP markers common for both breeding populations, including 5,598 SNPs located on the assembled parts of the genome (i.e. the 16 chromosomes).</p>	<p>Molecular data were obtained by GBS.</p> <p>Genome alignment: Bowtie2 software.</p> <p>Sequence data were processed Tassel GBS and Vcftools.</p> <p>We kept only Biallelic SNPs variants SNPs with 100% heterozygote genotypes were discarded. Individuals with more than 50% missing data were removed.</p> <p>Imputation of missing SNP data and phasing was carried out with Beagle 5.1.</p> <p>The Ne was estimated with the LD method of Waples and Do implemented in the NeEstimator 2.1 software.</p> <p>The computation was made separately in each population using the SNPs located on the genetic map.</p>	<p>Deli breeding population with effective population size (Ne) value of 3.0.</p> <p>La Mé breeding population with an effective population size value of 3.6.</p> <p>Deli breeding population has a 2.7-3.3 CIs value at a 95% confidence interval.</p> <p>La Mé breeding population has a 3.0-5.2 CIs value at a 95% confidence interval.</p>	<p>Ne values (<5) showed a significant degree of differentiation among the Deli and La Mé oil palm breeding populations.</p> <p>Considering the result, Deli needs more conservation breeding than La Mé for future use</p>

Appendix 11. Genetic map with 4,759 SNP markers on 16 linkage groups (LG). The y axis indicates the distances in centiMorgan (cM).



Appendix 11 (Continued). Genetic map with 4,759 SNP markers on 16 linkage groups (LG). The y axis indicates the distances in centiMorgan (cM).



PUBLISHED PAPERS

Genotyping by Sequencing for Plant Breeding- A Review



Essubalew Getachew Seyum^{1*}, Ngalle Hermine Bille¹, Joseph Martin Bell² and Wosene Gebreselassie²

^{1,2}Department of Plant Biology and Physiology, University of Yaoundé I, Cameroon

²Department of Horticulture and Plant Sciences, Jimma University College of Agriculture and Veterinary Medicine, Ethiopia

Submission: June 08, 2019; **Published:** August 30, 2019

***Corresponding author:** Seyum Esubalew Getachew, Department of Plant Biology and Physiology, Faculty of Sciences, University of Yaoundé I, Yaoundé, Cameroon

Department of Horticulture and Plant Sciences, Jimma University College of Agriculture and Veterinary Medicine, Jimma, Ethiopia

Abstract

Molecular plant breeding using DNA marker, or Marker-Assisted Selection (MAS), plays a pivotal role in a breeding program of crops to release a new variety within a short period of time compared to conventional method of plant breeding. Different types of marker have been used for breeding of plants and currently SNPs (single nucleotide polymorphisms) have become a reference type of DNA markers for plant breeding. Food production decline due to climate change, population growth, polyploid level and others result in a different level of DNA sequencing. Discovery of SNP without getting previous information about sequencing genome is called Genotyping by Sequencing (GBS). It is rapid, cost-effective and high throughput approach in next-generation sequencing. It is a new approach for implementing molecular tools in plant breeding. This paper briefly reviews the current status of Genotyping by Sequencing (GBS) and its application in plant breeding. It is efficiently applied in a wide range of plant breeding programs such as Genomic selection (GS), Genomic Diversity (GD), Genome-Wide Association (GWA), Linkage Analysis (LA), Marker discovery. It combines discovery of molecular marker and genome genotyping. This method has been developed and applied sequencing of multiplexed genomic samples. World economically most important crops, GBS used as tool for plant breeder to select them for better yield and quality. In spite of the above facts this method also have some limitation in its application.

Keywords: Molecular plant breeding; Single nucleotide polymorphism

Abbreviations: GBS: Genotyping-by-Sequencing; MAS: Marker-Assisted Selection; LA: Linkage Analysis; GWA: Genome-Wide Association; GD: Genomic Diversity; GS: Genomic selection; MAS: Marker-Assisted Selection

Introduction

Increasing production and productivity of crops for food and feed with the changing climate is one of the key slogans in our world in the 21st century [1]. Nowadays, agricultural productivity is becoming lower down due to biotic and abiotic stresses [2]. In this century world population will grow from 7 billion to 12.3 billion [3]. Reduction in crop production and productivity due to water scarcity, decreasing area and land degradation due to environmental change, pollution, occurrence of new pathogens and pests, and change in climate have major impact in food security of the world [2]. Improving production and productivity of major food, feed, and industrial crops in parallel alleviating food security problem plant breeding remains the main driving force [4]. To increase food production plant breeding will play a key role and breeders face an endless task in order to developing new crop varieties [5]. For this purpose, predicting population with the increasing climate change and considering both quantitative and

qualitative traits, yield stability should be a major focus of plant breeding.

Breeding of crops can be accomplished through two major approaches i.e., conventional and molecular. Variety development through the former approach requires continuous hybridization between distinct parents and selection over several generations. Long time (5-12 years) to develop crop variety, genotype by environmental interaction, low efficiency for complex and low heritable traits are the major limitations of this approach [4,6,7]. Applications of molecular biology tools that used to improve (develop) new cultivar is known as molecular plant breeding [8]. Unlike conventional method, this method used in DNA marker for selection of a given trait. This method helps to increase the efficiency, speed and precision of plant breeding in which it reduced cost and time [7,9].

Marker-Assisted Selection (MAS) selection process based on DNA marker for a given trait. It is a new discipline in the area of molecular breeding [10]. It is the method applied without phenotypic information in some individuals. It was started to solve the gaps in crop improvement program through conventional method [9]. It is tremendously useful in plant breeding and genetics. It is precondition for various biological applications such as mapping and tagging genes, segregation analysis, genetic diagnosis study, phylogenetic study etc. [11,12]. Selection of a trait and to know its association with a trait of interest in a target plants this method use DNA marker. It is more efficient for a character controlled by few Quantitative Trait Loci (QTLs) having major effect on trait expression. In contrary, this method is inferior over conventional breeding method in which a character controlled by a complex quantitative character [13-15]. A newly introduced approach in marker-assisted selection is known as genomic selection. This method uses high density genetic markers covering the whole genome in all Quantitative Trait Loci (QTL) and a genome linked with at least one marker. [16]. GS is used to estimate the genetic makeup an individual based on large set of markers distributed across the whole genome and selection was undertaken based on the relationship between training and validation sets, unlike the former it is not based on few markers [17-19] Genotyping-by-Sequencing (GBS) is newly introduced method and widely used range of crop improvement program in which it is used for detecting SNPs using high-throughput sequencing [20]. It is a modified RAD-seq based library preparation protocol for NGS [21]. The most important feature of this methods are reduced sample handling and fewer PCR purification steps, low cost, no reference sequence limits, no size fractionation and efficient barcoding technique [4] GBS was developed as a tool for genomic association studies and marker-assisted breeding. It is mainly works for species with large complex genomes and inimitable tool for genomics-assisted breeding in a wide range of plant species [22]. Presently, this technology has been used for whole genome sequencing and re-sequencing schemes in which the genomes of several specimens are sequenced to discover large numbers of Single Nucleotide Polymorphisms (SNPs) to discovering within-species diversity, constructing haplotype maps (blocks) and performing genome-wide association studies. Based on the above listed major problems and feeding the fastly growing population along with the problems it is essential to study modern breeding techniques. In the other hand around 7.4 million accessions of the world most economically important crops have no any non-model species it needs genotype sequencing [23]. Therefore, the objective of this paper is to review the role of genotyping by sequencing (GBS) in plant breeding and its application.

Molecular markers

Currently, molecular plant breeding has reached an advanced stage. For the last few decades different types of molecular markers have been used and develop [24]. The first DNA markers applied for plant genotyping were Restriction Fragment Length Polymorphism (RFLP) [25]. It is more suitable method in the

construction of genetic linkage maps. Despite its numerous advantages this approach becoming less applicable due to complicated hybridization, radioactivity, time consuming and limited number of available probes. Molecular plant breeding development resulted in the establishment numerous types of PCR-based markers mainly used in different crop improvement and research programs [24]

These PCR- based markers include Random Amplification of Polymorphic DNA (RAPD), Cleaved Amplified Polymorphic Sequences (CAPS) [26], Amplified Fragment Length Polymorphisms (AFLPs) [27], Simple Sequence Repeats (SSRs) [28], Sequence Characterized Amplified region (SCAR) [29] and Direct Amplification of Length Polymorphisms (DALP) [30]. Unlike RFLP, all these methods are relatively inexpensive and requires short period of time to undertake amplification and genome sequencing of a given populations [31]. Among all PCR based markers, the most applicable ones were Simple Sequence Repeats (SSR) and it was relatively inexpensive, abundant in plant genomes and more informative than bi-allelic markers [32]. In the year 1990s new techniques were developed by [33], for a given model plant species this method combines genome and Expressed Sequence Tags (ESTs). Identification of variations at the single base pair the development of Sanger sequencing highly accelerated the identification process [32]. The most recent DNA markers developed is Single Nucleotide Polymorphisms (SNPs) [34]. Plant genotyping through this technique has increased the potential to score variation in specific DNA targets. In addition, compared others it has small missing marker and also increases information on potentially millions of genome wide marker and their surrounding sequences sets in which it is the foundation of high-throughput genotyping [7,31,32]. Over the past 10 years, as compared to the earlier genotyping approaches, SNP-based marker techniques increased marker density, reducing cost of genotyping and requires less time for SNP discoveries [31]. The most common system in fluorescent detection of SNP-specific hybridization probes on PCR products are Taqman, Molecular Beacons and Invader [35]. In line with this, SNP-specific PCR primer extension products uses in homogeneous Mass-Extend (hME) assay. However, its output are read on a MALDI-TOF mass spectrophotometer [36]. Application all this method results around 100–1000s of SNPs per day. The current interest results an increasing demand for higher throughput, end-point fluorescent assays such as Taqman and Invader have been significantly enhanced by the use of array tape technology in place of 96, 384 or 1,536-well microtiter plates. This method reduced cost per assay and increasing throughput in a format [32].

Currently, there is enormously parallel array system enabled parallel scoring of up to hundreds and thousands of markers in plants genome. Depending on the application, assay simplicity, cost, throughput and accuracy, these ultra-high throughput technologies are used in wide range of researches. All these systems follow a similar pattern for DNA template preparation. The two most widely used array-based systems in plants genomic

are Golden Gate and Infinium assays and these arrays consist on multistep protocols based on Illumina's Bead Array/Bead Chip technology [37]. The former assay is allowing screening of many samples using a single multiplexed assay that can include as many as 3,072 SNPs. While, the latter assay provides considerably higher throughput, of up to four million SNPs from a single sample, or up to several hundred thousand on multiple samples in the same array. In Infinium, samples are incubated on bead chips where they anneal to locus-specific 50-mers covalently linked to beads. After hybridization, oligos are subject to allele-specific single-base extension; followed by fluorescent staining, signal amplification,

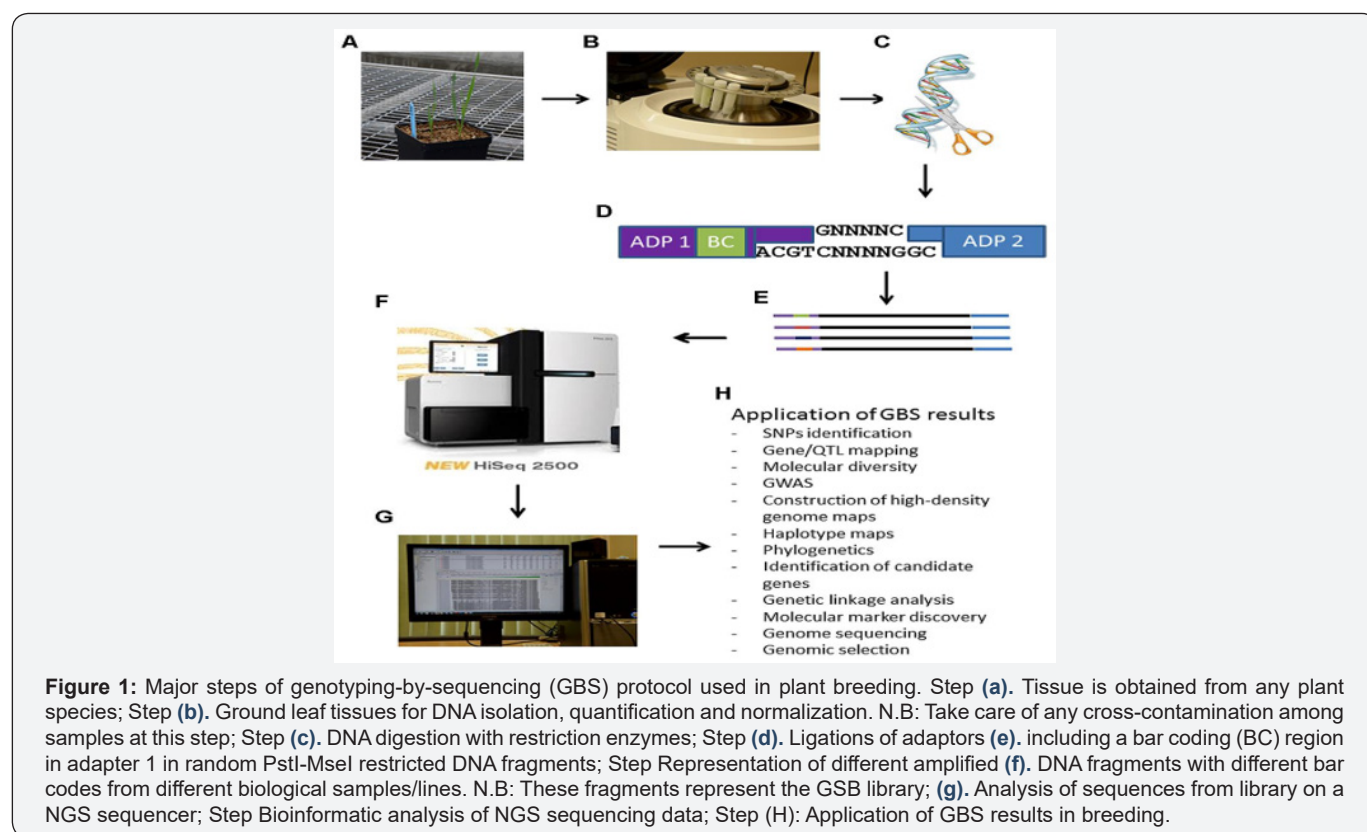
scanning in a dual-color channel reader, and analysis. The use of pre-made arrays reduces cost considerably although the actual number of markers derived from this array will be considerably lower, depending on the relationship to the reference and gene representation in the interrogated plants. Beckman Coulter's Genome Lab SNP stream is another method which allows the processing of up to three million genotypes in 384 samples/day/instrument (Table 1). Affymetrix Gene Chip system is most widely used method and it is not only detect hundreds of thousands of SNPs in a single array but, it can also be used for SNP discovery by Sequencing by Hybridization (SbH) [7,15,32,].

Table 1. Evaluation of representative NGS technologies.

No.	Sequencing Platform	Sequencing Chemistry	Detection Chemistry	RunTime ^a	Read Length (bp)	Reads per Run (million)	Throughput per Run (Gbp)
1	Roche 454 FLX Titanium	Sequencing by Synthesis	Light	23 hours	~800	~1	~0.7
2	Illumina MiSeq	Sequencing by Synthesis	Fluorescence	39 hours	2 × 250 b	~1	~8
3	Illumina HiSeq2500	Sequencing by Synthesis	Fluorescence	11 days (high output)/27 hours (rapid run)	2 × 100 b	~3,000	~600 (highoutput)/~120 (rapid run)
4	Life Technologies 5500xl	Sequencing by Ligation	Fluorescence	8 days	75 + 35 b	~5,000	~310
5	Ion Torrent PGM	Sequencing by Synthesis	pH	4 hours	100	1	~0.1

a Not including library construction; b Paired end read sequencing.

Genotyping-by-sequencing (GBS)



Genotyping by Sequencing (GBS) is the discovery of SNPs without prior knowledge about the genome sequence [20]. Nowadays, the advancement of NGS issue is the cost of DNA sequencing reduction to this end GBS is now feasible for large and complex genome species [21]. A thousand millions of SNPs can be detected in the large size lines that can be used for GWAS, GS, gd-study, linkage mapping, evolutionary studies and conservation and ecological genomics study [4,20,38]. It combines both discovery and genotyping of large populations genome applied in plant breeding even in the absence of a reference genome sequence. Its importance dramatically increases due to its cost-effective and unique tool for genomics-assisted breeding in a wide range of plant species [38, 39]. It is amenable to use on large numbers of individuals/lines due to library production procedure [4, 32]. Application of GBS technology in any plant species are summarized in Figure 1.

Genetic linkage map construction in a given test lines/ individuals GBS is more efficient and simpler in line with it combines with genome-independent imputation [21,40]. Originally the system used Ape KI protocol. Currently, modified to a two-enzyme namely PstI and MspI protocol, which reduced genome complexity and uniform library for sequencing than the

original protocol [39]. Now a days, GBS is applicable for different world most important economical food crops [41]. It is increase both SNPs call number and depth, allow an important reduction in per sample cost [4, 32].

Presently, it is an efficient approach for plant genotyping in NGS technologies is Reduction of Representation Library (RRL) [20]. The main component through this approach is cutting the entire genome with specific restriction enzyme(s) that reduce genome complexity for the organism of interest. Its results sequence dataset which can provide higher read coverage per locus while allowing higher level of multiplexing with uniquely bar-coded adapters for different samples [39]. The main limitation regarding RRL is that the important genomic regions may not be captured by GBS libraries when restriction sites are not available surrounding those regions. To overcome this problem, it is advisable to use multiple GBS libraries with different combinations of enzyme. Data depicted in Table 2 showed that different methods of GBS with their specific features for technical comparisons [6]. Different researches have been conducted in GBS for species with reference genomes and because of reference genome is available SNP genotyping becomes much easier than the other. Source.

Table 2: Representative GBS protocols published in peer-reviewed journals.

Method	Restriction enzyme	Insert size	Barcodes	Sequencing platform	Sequencing mode	Reference
RAD-seq (Restriction association DNA sequencing)	SbfI or EcoRI	Size-selection	~96	Illumina	Paired end	[42]
MSG (Multiplex shotgun genotyping) GBS (Genotype by sequencing)	MseI	Size-selection	~384	Illumina	Single end	[43]
	ApeKI	<350 bp	~384	Illumina	Paired end	[57]
Double-digested RAD-seq	EcoRI and MspI	Size-selection	~48	Illumina	Paired end	[44]
Double-digested GBS	PstI and MspI	<350 bp	~384	Illumina	Paired end	[22]
Ion Torrent GBS	PstI and MspI	<350 bp	~384	Ion Torrent	Paired end	[53]
SBG (Sequence-based genotyping)	EcoRI and MseI PstI and MseI	Size-selection	~32	Illumina	Paired-end	[46]
REST-seq (Restriction fragment sequencing)	TaqI and TruI	Size-selection	~305	Ion Torrent	Paired-end	[55]
Restriction enzyme sequence comparative analysis	MseI or NlaIII	Size-selection	~96	Illumina	Paired-end	[54]

In a GBS there are two different strategies which have been developed with the Ion PGM system for NGS [22]. Restriction enzyme digestion, in which no specific SNPs have been identified and ideal for discovering new markers for MAS programs. Multiplex enrichment PCR, in which a set of SNPs has been defined for a section of the genome. The first strategy works for all complex genome, which reduced its complexity by digesting the DNA with one or two selected restriction enzymes prior to the ligation of the adapters. The second approach designed to amplify the areas of interest by using PCR primers [40,42]. demonstrated that the first restriction site associated DNA sequencing or DNA (RAD) for high density SNP discovery and genotyping. It is a sequence-based marker and used to reduced-representation [32]. This barcoding system increased efficiency and relatively inexpensive.

Barcodes included sequences adapter and their locations, just upstream of the RE cut site in genomic DNA, eliminate the need for a second Illumina sequencing read. Unlike, RAD this system has modulation of barcode nucleotide composition and results in fewer length sequence phasing errors [9]. Substantially GBS becoming less complicated; generation of restriction fragments with appropriate adapters is more straight forward, single-well digestion of genomic DNA and adapter ligation results in reduced sample handling, there are fewer DNA purification steps, and fragments are not size selected as compared to the RAD method. Costs can be further reduced via shallow genome sampling tied with imputation of missing internal SNPs in haplotype blocks [20,40].

Libraries construction GBS mainly focuses on the reduction of genome complexity with the help of restriction enzymes [21]. Compared to the other approaches, GBS is simple, quick, extremely specific, highly reproducible, and may reach important regions of the genome that are inaccessible to sequence [40]. To get higher efficiency in GBS with a targeted of two or three-fold it needs the selection of appropriate REs, in order to avoid repetitive regions of genomes and lower copy regions [4,6]. This method tremendously simplifies computationally challenging alignment problems in species with high levels of genetic diversity [21].

Genotyping-by-sequencing (GBS) application in plant breeding

GBS is one of the most powerful tools in genome applications in the area of plant breeding. It is used to study GWAS, GS, gd-study, analysis of genetic linkage and marker discovery of non-model plants [22,40,43]. It is also an ideal platform for studying for a crop ranging from single gene to complex whole genome [4,40]. Generally, it is becoming an excellent tool for many applications and research questions in plant breeding and genetics for different food and industrial crops due to its flexibility and low cost [7,41]. According to it has been shown that this technique becoming valid tool to undertake genomic diversity studies. gd-GBS is new Illumina-based GBS protocol and it is unique from others. Compared to Roche 454 platform, this method yields more SNP genotype data with fewer missing observations. Genotyping

a diploid species, it uses of two restriction enzymes that used to reduce genome complexity, application of Illumina multiplexing indexes for barcoding and availability of a custom bioinformatics are the major features of gd-GBS. Like GBS, there are five major steps implemented gd-GBS (Figure 2&3). These are:

- a. Overall information about plant genetic diversity analysis;
- b. Specific genetic diversity project in mind to pursue;
- c. Plant materials prepared and ready to assay; and
- d. Access computing resources. The complete gd-GBS protocol, including the bioinformatics pipeline non-model plant genotyping np Geno, is provided in the online supporting materials [44].

As illustrated in Figure 2 & 3, GBS application in genetic diversity study (gd-GBS) involves five major steps:

- a. Sample preparation,
- b. Library assembly,
- c. Sequencing,
- d. SNP calling and
- e. Diversity analysis [44, 45].

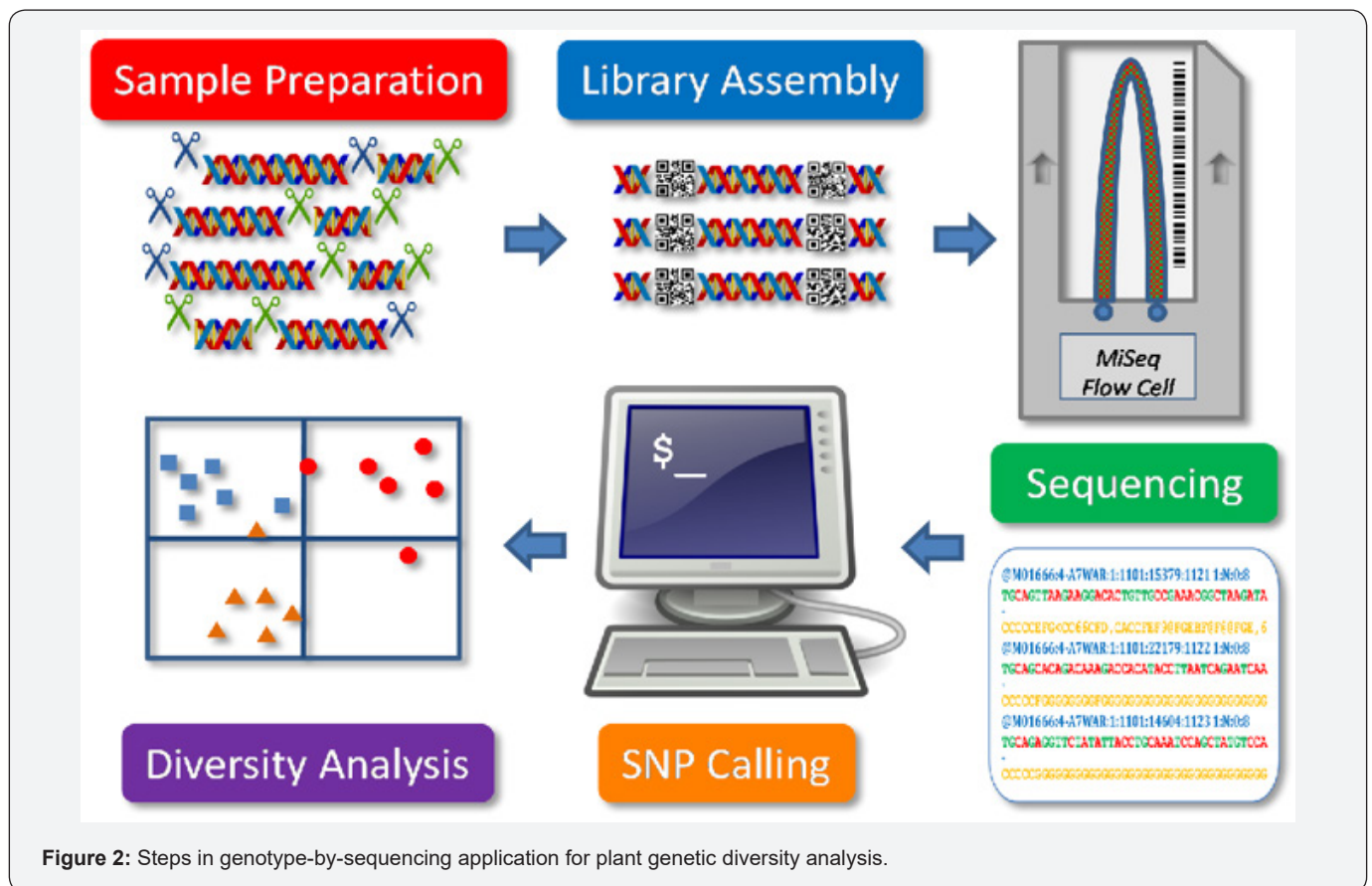


Figure 2: Steps in genotype-by-sequencing application for plant genetic diversity analysis.

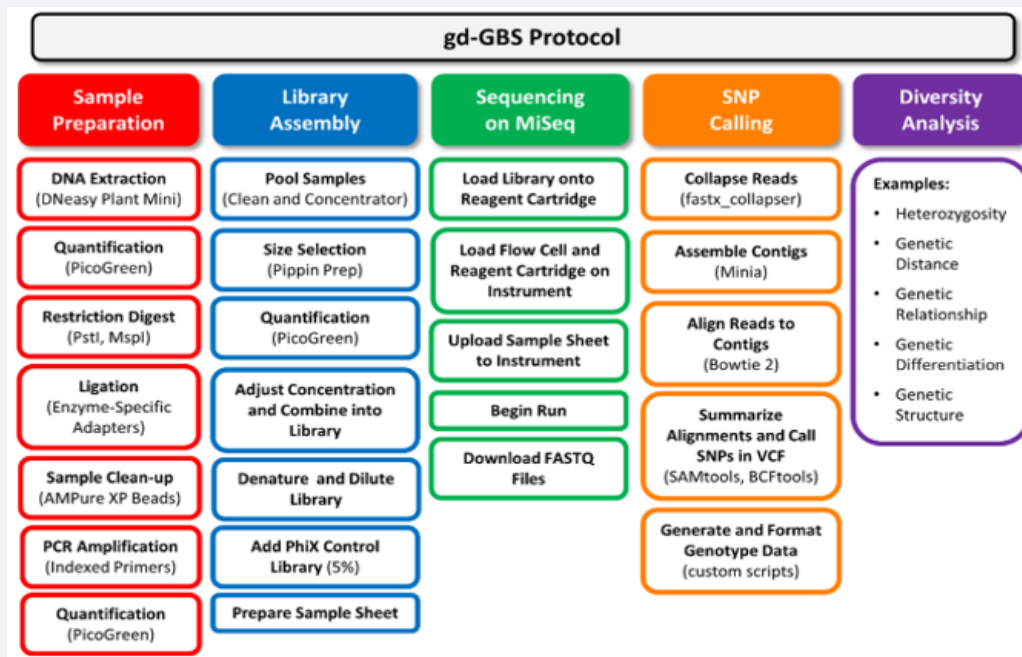


Figure 3: Flow chart of the genetic diversity-focused genotyping-by-sequencing (gd-GBS) protocol.

The above-mentioned steps may vary from one another depending on Restriction Endonuclease (RE) use and NGS platform and bioinformatics analysis for SNP genotyping to explore different objectives. To undertake genetic diversity analysis study the approach focuses on genome-wide sampling of many samples. Whereas, for genome-wide association studies, it emphasizes the accuracy of SNPs call rate with read depth to reveal genetic signals. Specifically, an informative genetic diversity analysis requires SNP data with large genome coverage, high genotyping accuracy, more balanced observation and less bias, which may be technically introduced from sequence mapping, heterozygote detection and data filtering [44]. In plant genetic diversity study analysis GBS approach have several major features. First, it combines the processes of marker discovery and genotyping, provides a rapid, high throughput and cost-effective tool for a genome-wide analysis of genetic diversity. Second, it requires no prior sequencing of the plant genome and provides direct genotyping of plants with complex genomes without prior SNP discovery. Third, and most importantly, it generates many genome-wide SNP data, allowing for better genome sampling. In general, this approach becoming more accessible for crop without model species [4, 6, 44].

To generate sufficient information and coverage in a GWAS it needs 100s of 1000s to millions of markers. However, the creation of NGS technologies greatly improved the resolution of marker [40]. Nowadays, GBS through the NGS has been used to sequence collections of recombinant inbred lines (RILs) to analyze and map various traits of interest for a specific breeding programs [32]. Cereals crop like maize, wheat, barley, sorghum, oat, rice, root and

tuber like potato, cassava and industrial crop like cotton have been reported to optimized by GBS for the efficient, low-cost and large scales of genome sequencing [32, 40, 46]. In maize a collection of 5,000 RILs have been sequenced using a restriction endonuclease-based approach and Illumina sequencing technology, which generated a total of 1.4 million SNPs and 200,000 indels [32, 40]. In maize an inclusive genotyping of 2,815 inbred accessions showed that 681,257 SNP markers are distributed across the entire genome, in which some SNPs are linked to the known candidate genes for kernel color, sweetness, and time of flowering [32, 47]. In soybean 31 genotypes with a set of 205,614 SNPs have been identified after resequencing giving valuable information for a soybean breeding programs. In potato [4, 40], 12.4 GB of high-quality sequence data and 129,156 sequence variants have been identified in breeding program of potato around 2.1 Mb were mapped to reference genome with a median average read depth of 636 per cultivar [32,40].

[48]reported that gd-GBS used the application of Roche 454 GS FLX Titanium technology with reduced genome representation and advanced bioinformatics tools to analyze 16 diverse barley landraces their genetic diversity and reported 2,578 contigs, and 3,980 SNPs, and confirmed a key geographical division in the cultivated barley gene pool [7]. The report from [49] showed that to access genetic diversity of species like switchgrass and they developed a SNP discovery pipeline based on a network approach called the Universal Network-Enabled Analysis Kit (UNEAK). Accordingly, 540 switchgrass plants sampled from 66 populations revealed informative patterns of genetic relationship with respect to ecotype, ploidy level, and geographic distribution to undertake

the diversity study. In addition, in mustard GBS protocol was used to analyze genetic diversity of 24 diverse yellow mustard accessions. The fining showed that 1.2 million sequence reads were generated, and 512 contigs and 828 SNPs were identified. Consequently, the genetic diversity study showed that yellow mustard SNPs revealed 26.1% of total variation over the landrace, cultivar, and breeding lines and 24.7% between yellow-seeded and black-seeded germplasm [7, 50].

In addition, sequencing of Arabidopsis in the whole genome shotgun sequencing on the Illumina platform a pool of 500 F2 plants generated by crossing a recessive Ethane Methyl Sulfonate (EMS)-induced Col-0 mutant characterized by slow growth and light green leaves, with a wild type Ler (Landsberg erecta) line. The result identifying high density SNP markers through GBS to construct genetic linkage maps which has a great value for numerous applications in plant breeding [7,51]. also reported that using a 384 plex GBS protocol to add 30,984 SNP markers to an Indica × japonica mapping population consisting of 176 rice recombinant inbred lines and mapped the recombined hot and cold spots and Quantitative Trait Loci (QTLs) for aluminum tolerance and leaf width. In bread wheat GBS was also applied resulting in the incorporation of 1000s of markers in the bread wheat map [22]. Identification of high resolution of SNP markers in barley and GBS mapping data were used to confirm that the semi-dwarfing gene (ari-e) is located on barley chromosome 5H [42, 49]. After the efficiency of multiplexed SNP genotyping for diversity, mapping and breeding applications were evaluated, and demonstrated that 384 plex SNP genotyping on the Bead Xpress platform is a robust and efficient method for marker genotyping and mapping in rice [32, 47].

The drawbacks of traditional method of plant breeding can be solved by MAS. With GBS, this is mainly achieved with the combination of molecular markers with phenotypic data to increase selection intensity and/or reduced selection interval on genotypic values [7]. Application of both applied and theoretical studies in genomic selection showed a great promise result to accelerate the rate new crop varieties (hybrid) development. GS through the GBS approach stands to be a major supplement to traditional crop improvement and it is a very important feature to move the genomics-assisted breeding into commercial crops [22]. GBS method on barley and wheat study without a reference genome provides a powerful method of developing high density markers by providing valuable tools for anchoring and ordering physical maps and whole genome shotgun sequence [40,47]. GBS approach also gives a very good promising result in cabbage, cauliflower and cotton without the reference genome identification and genetic diversity study. In Miscanthus the application of GBS is difficult due to ploidy level differences [47]. GBS approach also efficient to developed a catalog SNPs both within mapping population and among diverse African cassava varieties in which it allowing the improvements of MAS programs on disease resistance and nutrition in cassava [7].

Limitation of genotyping-by-sequencing

Despite the above listed advantages, the applications GBS have some potential drawbacks. In large, complex, polyploid genomes the difficulty getting aligned alleles in a single locus are the major challenges encountered by this method. Compared to others tools available to resolve the above problem GBS has a great potential. In addition, in hexaploid oat data analysis algorithms represent the main limiting factor to ascertain alleles at each single locus in a large polyploidy genome rather than GBS itself given sufficient depth of sequence is available [52]. reported that main weakness of GBS assay, when conducted at low coverage, is the amount of missing data. However, numerous imputation approaches are currently available, and yet more are presently in development, for a wide range of biological scenarios. Selecting appropriate imputation method and the probability of imputation success depends upon the biology of the study population. In the other hand, GBS genome complexity can be reduced by using restriction enzymes if applicable, in case of any mutation at the restriction site, the genomic DNA of this region is not available to be PCR amplified and consequently SNPs of this region will become unavailable and sometimes heterozygote gene may appear as homozygous. However, this drawback is not a problem only related with GBS rather it is shared by all the different methods involving reduction in genome complexity based on the utilization of restriction sites. GBS with two restriction enzymes have been overlooked to each other that the activity of MspI is inhibited in epigenetic studies. Therefore, developmental responses in plants may affect the SNP identification when using the enzyme MspI cannot be ignored but is likely reduced [7]. In addition, most of world food security crops (orphan crops) are neglected plant species and have not any known genomic sequence. An available reference genome can simplify the data analyses, but it is not essential in GBS for the above listed crops [7, 45].

Conclusion

World food security problem is one of the main agenda in the 21st century. To address this problem plant breeding is a main driving force [4] It can be accomplished by both conventional breeding and molecular breeding. However, the former approach has several limitations such as requiring a extended period of time to release high yielding variety. While, the later i.e., Marker-Assisted Selection (MAS) uses DNA markers and it is a new discipline in the area of 'molecular breeding' [4, 6, 47]. Currently, in a different crop improvement program a novel application in NGS that used to discovering and genotyping SNPs is known as Genotyping-by-sequencing (GBS). GBS has several advantages, including lower costs per samples, and relatively inexpensive to other whole genome genotyping platforms. Due to its use of high density of SNP markers, it is the most attractive approach to saturate mapping and breeding populations. Therefore, to attain the current problems in the area of plant breeding breeder's able to sequence and resequencing large crop genomes to this effect

it can establish high density genetic linkage maps from large size breeding populations. Even if it has the above listed advantages, it has also numerous biological and technical drawbacks. Among all the following points are considered as the major drawbacks in the application of GBS,

- a. Bias during PCR amplification and library construction,
- b. Lack of evenly covered regions of interest and within a given populations not all individuals are not sequenced very well,
- c. it requires continuous imputation for a missing data using both pedigree and parental information when available.

Future direction in GBS

Nowadays GBS has been reached an advanced stage but, some point regarding the limitation needs attention in the future. According [38] the following points should need more emphasis in the future regarding GBS New technical variation in GBS requires an advanced analytical tool for genomic data in which it can undergone genotyping large numbers of individuals and complete genotyping to the selected targets crops that are considered biologically, economically and socially relevant. Additionally, combination of GBS and RNA sequencing to find out SNPs in association with gene expression patters have a benefit to create a link between genomics, transcriptomics and proteomics. In general, this approach creates an opportunity to expand knowledge in the area of plant breeding and genetics research [53-57].

Acknowledgement

Special thanks goes for Dr David Cros form CIRAD for his very pertinent comment and suggestion which was very helpful in improving the manuscript.

References

1. Spiertz JHJ, Ewert F (2009) Crop production and resource use to meet the growing demand for food, feed and fuel: opportunities and constraints. *NJAS - Wagening J Life Sci* 56(4): 281-300.
2. Collard, BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Trans R Soc B Biol Sci* 363(1491): 557-572.
3. Kim C, Guo H, Kong W, Chandnani R, Shuang LS, et al. (2016a). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Sci* 242: 14-22.
4. Bhat JA, Ali S, Salgotra RK, Mir ZA, Dutta S, et al. (2016) Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. *Front. Genet* 7: 221.
5. Evans LT, (1997) Adapting and improving crops: the endless task. *Philos. Trans R Soc B Biol Sci* 352: 901-906.
6. Kim C, Guo H, Kong W, Chandnani R, Shuang LS, Paterson AH (2016b). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Sci* 242: 14-22.
7. He J, Zhao X, Laroche A, Lu ZX, Liu H, et al. (2014a) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front Plant Sci* 5: 484.
8. Xu Y (2010) Molecular plant breeding. *Mol Plant Breed* 1-734.
9. Barabaschi D, Tondelli A, Desiderio F, Volante A, Vaccino P, et al. (2016) Next generation breeding. *Plant Sci* 242: 3-13.
10. Collard BCY, Mackill DJ (2008b) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Philos Trans R Soc Lond B Biol Sci* 363(1491): 557-572.
11. Dale Young N (1999) A cautiously optimistic vision for marker-assisted breeding. *Molecular Breeding* 5(6): 505-510.
12. Mohan M, Nair S, Bhagwat A, Krishna TG, Yano M, et al. (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding* 3(2): 87-103.
13. L Heffner E, J Lorenz A, Jannink JL, Sorrells M, (2010) Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. *Crop Sci - CROP SCI* 50(5): 1681-1690.
14. Varshney RK, Mohan SM, Gaur PM, Gangarao NV, Pandey MK, et al. (2013) Achievements and prospects of genomics-assisted breeding in three legume crops of the semi-arid tropics. *Biotechnol Adv* 31(8): 1120-1134.
15. Zhong S, Dekkers JC, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a Barley case study. *Genetics* 182(1): 355-364.
16. Goddard ME, Hayes BJ (2007) Genomic selection. *J Anim Breed Genet* Z 124(6): 323-330.
17. 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.
18. Kwong QB, Ong AL, Teh CK, Chew FT, Tammi M, et al. (2017) Genomic Selection in Commercial Perennial Crops: Applicability and Improvement in Oil Palm (*Elaeis guineensis Jacq*). *Sci Rep* 7(1): 2872.
19. Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci* 19(9): 592-601.
20. Scheben A, Batley J, Edwards D (2017) Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnol J* 15(2): 149-161.
21. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE* 6(5): e19379.
22. Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLOS ONE* 7(2): e32253.
23. Wambugu PW, Ndjiondjop MN, Henry RJ (2018) Role of genomics in promoting the utilization of plant genetic resources in genebanks. *Brief Funct Genomics* 17(3): 198-206.
24. Grover A, Sharma PC (2014) Development and use of molecular markers: Past and present. *Crit Rev Biotechnol* 36(2): 290-302.
25. Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121(1): 185-199.
26. Jarvis P, Lister C, Szabo V, Dean C (1994) Integration of CAPS markers into the RFLP map generated using recombinant inbred lines of *Arabidopsis thaliana*. *Plant Mol Biol* 24(4): 685-687.
27. Vos P, Hogers R, Bleeker M, Reijmans M, van de Lee T (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* 23(21): 4407-4414.
28. Salimath SS, de Oliveira AC, Godwin ID, Bennetzen JL (1995) Assessment of genome origins and genetic diversity in the genus *Elaeusine* with DNA markers. *Genome* 38(4): 757-763.

29. Jiang C, Sink KC (1997) RAPD and SCAR markers linked to the sex expression locus M in *asparagus*. *Euphytica* 94(3): 329-333.
30. Desmarais E, Lanneluc I, Lagnel J (1998) Direct amplification of length polymorphisms (DALP), or how to get and characterize new genetic markers in many species. *Nucleic Acids Res* 26(6): 1458-1465.
31. Amom T, Nongdam, P (2017) The Use of Molecular Marker Methods in Plants: A Review.
32. Deschamps S, Llaca V, May GD (2012) Genotyping-by-Sequencing in Plants. *Biology* 1(3): 460-483.
33. Rounsley SD1, Glodek A, Sutton G, Adams MD, Somerville CR, et al. (1996) The Construction of Arabidopsis Expressed Sequence Tag Assemblies A New Resource to Facilitate Gene Identification. *Plant Physiol* 112(3): 1177-1183.
34. Wang DG, Fan JB, Siao CJ, Berno A, Young P, et al. (1998) Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome. *Science* 280(5366): 1077-1082.
35. Takatsu K, Yokomaku T, Kurata S, Kanagawa T (2004) A new approach to SNP genotyping with fluorescently labeled mononucleotides. *Nucleic Acids Res.* 32(7): e60.
36. Elvidge GP, Glennly L, Appelhoff RJ, Ratcliffe PJ, Ragoussis J, et al. (2006) Concordant Regulation of Gene Expression by Hypoxia and 2-Oxoglutarate-dependent Dioxygenase Inhibition THE ROLE OF HIF-1 α , HIF-2 α , AND OTHER PATHWAYS. *J. Biol. Chem.* 281(22): 15215-15226.
37. Gupta PK, Rustgi S (2018) Array-Based High-Throughput DNA Markers and Genotyping Platforms for Cereal Genetics and Genomics.
38. Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Mol Ecol* 22(11): 2841-2847.
39. eHe J, eZhao X, eLaroche A, eLu Z-X, eLiu H, eLi Z (2014) Genotyping by sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front Plant Sci* 5: 484.
40. He J, Zhao X, Laroche A, Lu Z-X, Liu H, Li Z (2014c). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci* 5: 484.
41. Poland JA, Rife TW (2012a) Genotyping-by-Sequencing for Plant Breeding and Genetics. *Plant Genome* 5(3): 92-102.
42. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS One* 3(10): e3376.
43. Andolfatto P, Davison D, Erezylmaz D, Hu TT, Mast J, et al. (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res* 21 (4): 610-617.
44. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE* 7(5): e37135.
45. Poland JA, Rife TW (2012b) Genotyping-by-Sequencing for Plant Breeding and Genetics. *Plant Genome J* 5: 92.
46. van Poecke RM, Maccaferri M, Tang J, Truong HT, Janssen A, et al. (2013) Sequence-based SNP genotyping in durum wheat. *Plant Biotechnol. J* 11(7): 809-817.
47. He J, Zhao X, Laroche A, Lu Z-X, Liu H (2014b) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front Plant Sci* 5.
48. Fu YB, Peterson GW (2011) Genetic diversity analysis with 454 pyrosequencing and genomic reduction confirmed the eastern and western division in the cultivated barley gene pool. *Plant Genome* 4: 226-237.
49. Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, et al. (2013) Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *PLOS Genet* 9 e1003215.
50. Fu YB, Cheng B, Peterson GW (2014) Genetic diversity analysis of yellow mustard (*Sinapis alba* L.) germplasm based on genotyping by sequencing. *Genet. Resour. Crop Evol* 61: 579-594.
51. Spindel J, Wright M, Chen C, Cobb J, Gage J, et al. (2013) Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. *Theor Appl Genet* 126 2699-2716.
52. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, et al. (2014) TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *Plos One* 9(2): e90346.
53. Mascher M, Wu S, Amand PS, Stein N, Poland J (2013) Application of Genotyping-by-Sequencing on Semiconductor Sequencing Platforms: A Comparison of Genetic and Reference-Based Marker Ordering in Barley. *PLOS ONE* 8(10): e76925.
54. Monson-Miller J, Sanchez-Mendez DC, Fass J, Henry IM, Tai TH, et al. (2012) Reference genome-independent assessment of mutation density using restriction enzyme-phased sequencing. *BMC Genomics* 13: 72.
55. Stolle E, Moritz RFA (2013) RESTseq - Efficient Benchtop Population Genomics with RESTriction Fragment SEQuencing. *PLOS ONE* 8: e63960.
56. Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res* 18(22): 6531-6535.
57. Yu J, Hu S, Wang J, Wong GK, Li S, et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* 296(5565): 79-92.



This work is licensed under Creative Commons Attribution 4.0 License
DOI: [10.19080/AIBM.2019.14.555891](https://doi.org/10.19080/AIBM.2019.14.555891)

**Your next submission with Juniper Publishers
will reach you the below assets**

- Quality Editorial service
- Swift Peer Review
- Reprints availability
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats
(Pdf, E-pub, Full Text, Audio)
- Unceasing customer service

Track the below URL for one-step submission
<https://juniperpublishers.com/online-submission.php>

Genome Mapping to Enhance Efficient Marker-Assisted Selection and Breeding of the Oil Palm (*Elaeis guineensis* Jacq.)

Essubalew Getachew Seyum^{1,2,3*}, Ngalle Hermine Bille¹, Wosene Gebreselassie Abteaw³, Godswill Ntsomboh-Ntsefong^{1,4}, Joseph Martin Bell¹

¹Department of Plant Biology and Physiology, Faculty of Sciences, University of Yaounde I, Yaounde, Cameroon

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

³Department of Horticulture and Plant Sciences, Jimma University College of Agriculture and Veterinary Medicine, Jimma, Ethiopia

⁴Institute of Agricultural Research for Development, Yaounde, Cameroon

Email: *g.essu2011@gmail.com, *getachew.essubalew@cirad.fr, hbille2014@gmail.com, wosish@yahoo.com, ntsomboh@yahoo.fr, josmarbell@yahoo.fr

How to cite this paper: Seyum, E.G., Bille, N.H., Abteaw, W.G., Ntsomboh-Ntsefong, G. and Bell, J.M. (2021) Genome Mapping to Enhance Efficient Marker-Assisted Selection and Breeding of the Oil Palm (*Elaeis guineensis* Jacq.). *Advances in Bioscience and Biotechnology*, 12, 407-425. <https://doi.org/10.4236/abb.2021.1212026>

Received: September 29, 2021

Accepted: December 10, 2021

Published: December 13, 2021

Copyright © 2021 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0). <http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

The oil palm (*Elaeis guineensis* Jacq.) is one of the major cultivated crops among the economically important palm species. It is cultivated mainly for its edible oil. For a perennial crop like oil palm, the use of Marker Assisted Selection (MAS) techniques helps to reduce the breeding cycle and improve the economic products. Genetic and physical maps are important for sequencing experiments since they show the exact positions of genes and other distinctive features in the chromosomal DNA. This review focuses on the role of genome mapping in oil palm breeding. It assesses the role of genome mapping in oil palm breeding and discusses the major factors affecting such mapping. Generating a high-density map governed by several factors, for instance, marker type, marker density, number of mapped population, and software used are the major issues treated. The general conclusion is that genome mapping is pivotal in the construction of a genetic linkage map. It helps to detect QTL and identify genes that control quantitative traits in oil palm. In perspective, the use of high-density molecular markers with a large number of markers, a large number mapping population, and up-to-date software is necessary for oil palm genome mapping.

Keywords

Genome Mapping, Crop Improvement, Marker Assisted Selection,

1. Introduction

The oil palm (*Elaeis guineensis* Jacq., Areaceae) originated from West Africa [1] [2]. It is a diploid ($2n = 2x = 32$), perennial monocot plant, and the most productive oil-producing crop in the world. It is mainly cultivated in humid tropical zones of the world [3] [4]. It is naturally cross-pollinated, monoecious, and allogamous. The economic life span of the oil palm is about 30 years [5] [6].

According to the United States Department of Agriculture (USDA) [7], the total world vegetable oil production as of 2020/2021 was 72,769 MT, led by soybean oil (362.64 MT), followed by palm oil (75.19 MT), and rapeseed oil (69.17 MT) and sunflower seed (49.66 MT). Oil palm produces on average 4 metric tons of oil per hectare every year [8] [9]; this is approximately 10 times higher than soybean. Palm oil falls into two major applications: the food industry (with over 80% of the market) and the rest for the chemical industry for formulation of paints, inks, resins, varnishes, plasticizers, biodiesel production, etc. [3] [10].

Despite its wide adaptation and importance, oil palm production and productivity are generally far from their potential due to several biotic and abiotic constraints. Climate change, land, and a labor shortage are major factors that hinder yield and palm oil quality across the world [11]. Moreover, breeding of oil palm is made difficult because of the perennial nature of the crop that limits the rate of increasing palm oil yield and quality. Equally, Herrero *et al.* [12] reported that breeding of oil palm is applicable through the use of the conventional method, which needs more space and time for selecting promising crosses, mainly when increasing parental biodiversity. To alleviate this problem and improve oil palm yield and quality, breeders need to implement molecular techniques of oil palm breeding.

In marker-assisted selection (MAS), the use of a molecular marker with quantitative trait loci (QTL) helps in phenotypic screening to address the limitations of traditional breeding methods. The accuracy and efficiency of selection are improved by MAS [13]. The method brings a remarkable result mainly for traits with low to moderate heritability, which is difficult to achieve by the traditional breeding method. In most cases, MAS breeding requires knowledge about the distribution of QTL for the targeted trait inside the genome. In many crop species including the oil palm, MAS has been instrumental in the genetic improvement of several agronomic traits [13]. In oil palm, the use of MAS studies has been discussed since the 1990s [14]. In whole genome sequencing research, linkage maps, molecular markers, and QTL maps are crucial for MAS. In several crop species, linkage maps, a large number of DNA markers, and identification of QTL for major traits have been developed [15].

Genetic linkage maps express the actual inheritance of loci into offspring based on the patterns of recombination during meiosis [16]. In the oil palm, different genetic linkage maps from numerous families of oil palm have been constructed with remarkable results based on MAS breeding studies. Some genetic

maps have been constructed for the oil palm using amplified fragment length polymorphism (AFLP) [17] [18], restriction fragment length polymorphism (RFLP) [14] [18], random amplified polymorphic DNA (RAPD) [19], simple sequence repeats (SSR) [20] [21] [22] [23], and single nucleotide polymorphism (SNP) [12] [24] [25] [26] [27]. Still, with the oil palm, numerous quantitative trait loci (QTL) mapping reports have revealed the existence of major-effect QTLs for many traits [12] [15] [18] [19] [20] [23] [27] [28] [29] [30]. The objective of this review was to assess and highlight the role of genome mapping in oil palm breeding and discuss the major factors affecting genome mapping.

2. Types of Genome Maps

Several types of maps exist such as cytogenetic map, physical map, and genetic map.

2.1. Cytogenetic Map

This is the visual appearance of a chromosome when stained and/or labeled under a microscope [31] [32]. The units for the cytogenetic map are a fraction of a chromosomal arm or centiMcClintocks (cMc) (**Figure 1**) [33]. This is obtained by visualizing distinct regions marked by light and dark bands which give each of the chromosomes a unique appearance. The map shows the positions of chromosomes in the bands, *i.e.* bandmap (genome deletion panel) [33] [34]. Hozier and Davis [35] showed that the integration of this method of mapping with other molecular genetic mapping methods allows the study and mapping of different mammalian genomes. Azhaguvel *et al.* [36] reported that this type of map is the earliest that has been used for mapping fruit fly and the corn crop. Shah [34] on their part showed that the application of the cytogenetic mapping method is advantageous to study genome analysis, chromosome mapping, and analysis of somaclonal variations in tissue culture.

2.2. Physical Map

Physical mapping reflects the actual physical distance in base pairs (bp) or multiples thereof (for example, kilobases (kb) *i.e.* $\text{bp} \times 1000$) between molecular markers (**Figure 1**) [37] [38]. Such maps are increasingly being used to understand the molecular insights of genes and their evolution [36]. According to O'Rourke [16], physical maps provide an effective tool to isolate and study genes: where they are, what they do, and how they interrelate? A better understanding of these maps allows the location of the marker in the chromosome with the centromere and telomeres and permits the detection of some mutation phenomena such as insertions, deletions, and translocations [39]. Due to the current advancement in sequence technology, there is a constant increase of interest in these maps mainly because of the difficulty of assembling large fractionated genomes without a good physical map [40]. Dixit [33] stated that, unlike genetic maps, the construction of a physical map requires molecular biology techniques; indeed, it represents the entire genome as a set of overlaying cloned

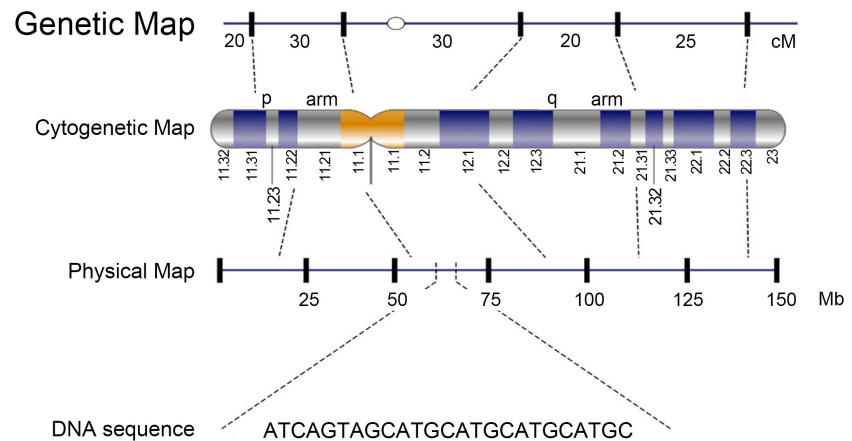


Figure 1. Illustration of physical, genetic map, and position of markers in the chromosome.

DNA fragments that make up a genome and this is ordered with respect to a reference map (such as a genetic map). In the same light, Deonier [38] reported the usefulness of the construction of wide-ranging physical maps for studying the characteristics of both sequenced and unsequenced genomes. These maps are used in the genomic study of oil palm, in addition to genetic maps. For instance, Herrero *et al.* [12] reported that this map can be used for the analysis of quantitative trait locus for the traits of interest, generally for crop and specifically oil palm breeding.

2.3. Genetic Map

Genetic maps show the positions of genes and other related sequence features like DNA markers in the genome, based on how often the genes or markers are inherited together [31] [32] [37]. It shows the map distance, in cM, which separates any two loci, and the position of these loci relative to all other mapped loci (Figure 1) [41]. It is constructed based on the meiotic recombination between homologous chromosomes [36]. The use of high-resolution genetic maps is very pertinent since they determine the relationship between breeding and genome sequencing. Likewise, O'Rourke [16] outlined that in the genetic and genomic studies of plants or animals, the application of high-density genetic maps full of polymorphic markers is a key for marker-assisted selection. Conversely, Li [42] noted that the construction of high-density and high-resolution genetic maps is vital to the structural and functional understanding of the genome and genes of interest through linkage analysis. Dixit [33] showed that a high resolution of the genetic map is determined by several factors such as the number of crossovers in a plant species that have been scored or a large number of progeny in humans. They also stated that the type of molecular markers used also has an impact on the resolution of the map.

3. Genetic vs. Physical Maps

Genetic and physical maps demonstrate the arrangement of genes and DNA

markers on a chromosome, respectively. Details of their differences are presented in **Table 1**. The genetic map is also known as a linkage map. It describes genes or loci within a chromosome based on recombination rate [36]. This mapping concept was first developed by Sturtevant [43], established by linearly placing five sex-linked genes on the Y chromosome of fruit fly (*Drosophila melanogaster*). It provides an approximate distance between loci in the genome in terms of recombination rate and its determination is based on the number of crossovers [33]. O'Rourke [16] showed that crossing over frequency between genes or DNA markers is proportional to the chromosomal distance between them. For instance, the more closer the genes, the fewer crossover frequencies, and vice-versa. It allows the establishment of linkage associations between genes or DNA markers and it is a baseline to establish the physical mapping, thereby opening a door for map-based genome isolation. This map does not allow the study of particular chromosomes in the genome, rather a set of polymorphic genetic marker loci or genes [36]. In this map, the genetic distance between two molecular markers is computed based on the number of recombination events without precision on the actual physical distance [33]. To address the above problem, physical mapping has to be performed.

A physical map has a linearly ordered set of molecular markers (DNA fragments) surrounding the whole genome or a particular genomic region of interest [31] [32]. Azhaguvel [36] classified this map into two types. The first is the macro-restriction map which gives information regarding the DNA fragments at the chromosome level. The second is known as the ordered clone map consisting of an overlapping collection of cloned DNA fragments, such as in yeast and bacterial artificial chromosome (YAC). It determines the actual distance of DNA markers on the chromosomes in base pairs [33]. The genetic-physical map ratio varies significantly from one chromosome region to the other. It is mainly dependent on the nature of the chromosome and the frequency of recombination in that region [36]. For example, the estimated genetic to physical distance ratio of oil palm range from 68.44 Mb/cM to 21.37 Mb/cM [25].

Table 1. Differences between genetic and physical maps.

Genetic map	Physical map
It is the calculated map distance based on the crossover percent between two linked genes	The actual physical distance between linked genes
This map distance highly varies as the frequency of crossing over varies in a different segment of chromosomes and it is only a predicted value	The physical distance of linked genes bears no direct relationship to the map distance calculated based on crossover percentages
The distance measured in Map unit or centiMorgan	The distance measured in base pairs (bp, Kbp, Mbp)
Linear order is identical as in the physical map	Linear order is identical as in the genetic map
The relative distance between two genes	The exact location of genes in the chromosomes

The need for physical-genetic maps has increased steadily in the past decade. Since then these two maps are used fully to study gene cloning and whole-genome and specific genome region DNA sequencing [31] [32]. A genetic map constructed to identify the target gene and closely linked DNA markers were used to filter a large set of the library used to construct the physical map [36]. Then, the newly produced DNA markers were used to identify the clones for genetic fine mapping. O'Rourke [16] reviewed the correlation of genetic and physical maps and revealed that physical maps consist of ordered library pieces of DNA covering entire genomes or chromosomes; the genetic map was constructed based on the recombination analysis of molecular markers, with the main target to identify the cloning genes. Physical and genetic map integration is used to identify the genomic region that has a high recombination hot spot region with repressed recombination [44]. Azhaguvel [36] noted that such integration reveals all about the genome sequences. This opens a new door to develop DNA markers, identify genes, quantitative trait loci (QTLs), expressed sequence tags (ESTs), regulatory sequences, and repeat elements.

4. A Molecular Marker Used for Mapping in Oil Palm Populations

Molecular markers are widely used nowadays in various plant breeding programs to track loci and genomic regions [36]. Identification of major genes controlling quantitative traits in crop plant genomes is possible with molecular markers. To this end, genetic mapping techniques are used to retrieve and locate important genes and genomic information responsible for a particular trait [31] [32]. Several genetic maps have been established for a wide range of plant species using various molecular marker systems such as RFLP [45], RAPD [46], simple sequence repeat (SSR), or microsatellite [47], sequence-tagged sites (STS) [48], AFLP [49], single-nucleotide polymorphism (SNP) [50], sequence-characterized amplified region (SCAR) [51], and cleaved amplified polymorphic sequences (CAPS) [52]. Depending on different purposes for gene mapping, each of the molecular markers has its pros and cons (Table 2). However, RFLP, RAPD, SSR, and AFLP markers are most commonly used in plant species for genetic mapping [16].

In the oil palm, different molecular marker types have been used for the construction of genetic linkage maps (Table 3). They include RFLPs [14] [18] [29] [30], RAPDs [19], AFLPs [18] [19] [22] [23] [30], SSRs [3] [15] [18] [19] [20] [21] [23] [30] [53] [54] [55], and SNPs [12] [15] [24] [25] [26] [27] [28] [53] [54] [55]. In this paper and for the first time, we review the primary results of oil palm genome mapping as summarized in Table 3. This table shows an outline of the major studies on oil palm genome mapping with their different features.

5. Oil Palm Genome Mapping

Genetic linkage maps reflect the actual inheritance of loci from parents to their offspring based on the patterns of recombination during meiosis. In oil palm for

Table 2. Evaluation of the most widely used molecular marker systems.

Marker Name	PCR-based	Polymorphism (abundance)	Dominance	Reproducibility	Automation	Running cost
RFLP	No	Low/medium	Codominant	High	Low	High
RAPD	Yes	Medium/high	Dominant	Low	Medium	Low
SCARS/CAPS	Yes	High	Codominant	High	Medium	Medium
AFLP	Yes	High	Dominant	High	Medium/high	Medium
SSR	Yes	High	Codominant	High	Medium/high	Low
ISSR	Yes	High	Dominant	High	Medium/high	Low
STS	Yes	High	Codominant/ dominant	High	Medium/high	Low
SRAP/EST	Yes	Medium	Codominant	High	Medium	Low
IRAP/REMAP	Yes	High	Codominant	High	Medium/high	Low
SNP	Yes	Extremely high	Codominant/ dominant	High	High	Low

Table 3. Summary of linkage map constructed in oil palm.

No.	Year	Type of markers	No of Markers	Map depth (cM)	No of LGs	Software use to construct LGs	References
1	1997	RFLP	97	860	24	MAPMAKER 2.0	[14]
2	2000	RAPD	48	399.7 - 449.3	12 - 15	MAPMAKER 2.0	[19]
3	2001	RFLP	153	852	22	JoinMap 2.0	[29]
4	2005	MS and AFLP	255 + 688	1743	16	JoinMap ver. 3.0	[17]
5	2009	AFLP, RFLP, MS	252	1815	21	Joinmap ver. 4.0.	[18]
6	2010	SSR	251	1479	16	JoinMap v. 3.0	[20]
7	2011	AFLP	331	2274.5	16	MAPRF7	[22]
8	2013	AFLP, RFLP, SSR	148	798.0	23	JoinMap 4.0	[30]
9	2013	SSR	362	1845.0	16	JoinMap v.4.0	[3]
10	2014	SSR, Genes, SNP	190	1233.0	31	JoinMap v.4.0	[54]
11	2014	SSR, AFLP	423	1931	16	JoinMap v 3.0	[23]
12	2014	SSR, SNP	1331	1867	16	JoinMap v 4.1	[55]
13	2015	SSR, SNP	480	1565.6	16	JoinMap v 3.0	[15]
14	2015	SNP	1085	1429.6	16	JoinMap v 3.0	[27]
15	2015	SSR	281	1935.0	16	CRIMAP	[21]
16	2018	SNP, SSR	10,023	2398.2	16	Lep-MAP v 2	[53]
17	2018	SNP, DArTseq	1399 - 1466	1873.7 - 1720.6	16	JoinMap v 4.1	[24]
18	2018	SNP	2413	1161.89	16	JoinMap v 4.1	[28]
19	2019	SNPs	27,890	1151.7	16	Lep-MAP3	[25]
20	2020	SPET	3501	1,370	16	Lep-MAP3	[12]
21	2020	SNPs	11,421	1151.70 - 1268.26	17 - 24	Lep-MAP3	[26]
22.	2022	SNPs	-	-	16	Lep-MAP3	Essubalew <i>et al.</i> 2022 (under review)

the last 20 years, several linkage maps have been constructed and used to detect different vegetative, yield, and yield components and palm oil quality traits [25]. In the same light, markers like RFLPs, AFLPs, SSRs, and SNPs are widely used to construct genetic linkage maps in oil palm. Recently, restriction associated DNA tagging (RAD), double digestion RAD (ddRAD), single primer enrichment technology (SPET) have been recognized for producing a large number of SNPs with remarkable maps [12].

In oil palm, the first genetic linkage map constructed based on RFLP markers from genomic libraries was published in 1997. This map which considers 97 RFLP markers (84 probes) mapped a selfed *guineensis* cross (*tenera* x *tenera*) with a total genetic distance of 860 cM, and produced a total of 24 linkage groups using a LOD score of 4 and recombination fraction of 0.4. According to the study [14], more than 95% of the markers could be linked to at least one other marker, suggesting that good genome coverage helps to detect the position of the shell thickness gene (*Sh*) at a distance of 9.8cM on group 10. From their result, Mayes *et al.* [14] concluded that this map helps to enable the mapping of the gene responsible for controlling major commercial oil palm traits. Likewise, Rance [29] also used 153 RFLP markers to construct a genetic linkage map of 84 self-fertilization F2 oil palm populations used to detect major genes influencing shell thickness. The result confirms that QTL mapping helps to detect genes that influence a large proportion of the total phenotypic variance in a large and small population.

Further, RAPD is another marker used to construct a genetic linkage map in the oil palm. The first RAPD marker map was developed by Moretzsohn [19] to develop a pseudo-testcross mapping strategy in combination with the RAPD assay. This was meant to construct genetic linkage maps of different fruit types (shell thickness) of F₁ *tenera* (*sh+* *sh-*) x *pisifera* (*sh-* *sh-*) progeny populations. The map used a total of 48 RAPD markers, and 308 F1 progeny populations, and produced a total of 12 linkage groups with a map distance ranging from 399.7 - 449.3 cM at a LOD score of 5.0 and by considering the projected *Elaeis* total map distances and genome sizes, physical and genetic distances relationships were established (1.06 Mbp/1 cM and 1.09 Mbp/1 cM, for *tenera* and *pisifera*, respectively). They also obtained limited genome coverage with the two maps (28.0%, for *tenera* and 25.6%, for *pisifera*). This result depicted the importance of RAPD markers used for genetic linkage mapping markers closer to the *sh+* locus, helped to detect the gene responsible for shell thickness, and gave a step forward for MAS for shell thickness in the oil palm.

AFLP is another pronounced marker used to construct a genetic linkage map in the oil palm. The first AFLP based genetic map in oil palm was developed by Billotte *et al.* [17] involving a cross between a thin-shelled *E. guineensis* (*tenera*) palm and a thick-shelled *E. guineensis* (*dura*) palm with the main goal of mapping to detect the presence and absence of the gene responsible for shell in the oil palm fruit. For this purpose, they used a total of 944 (255 SSRs, 688 AFLPs,

allele *Sh*-) markers with a map length of 1743 cM and with an average of one marker every 1.8 cM and LOD score 3.0, producing a total of 16 linkage groups. The lengths of the linkage groups varied between 59 cM and 192 cM. This map was the first linkage map for the oil palm to have 16 independent linkage groups corresponding to the haploid chromosome number of 16 in the oil palm. Their findings on high-density linkage maps could step forward research for QTLs and physical mapping in the *E. guineensis* species. Besides, they also reported that SSRs marker had better mapping resolution compared to that of AFLPs. This is because high-density markers like SSRs have higher recombination rates than low-density markers like AFLPs. From the result, they observed that SSRs markers are well distributed along the genome than AFLPs markers. Conversely, Singh [18] reported an interspecific cross involving Colombian *Elaeis oleifera* (UP1026) and a Nigerian *E. guineensis* (T128) and a total of 118 palms from this interspecific cross were used to detect quantitative trait loci (QTLs) controlling oil quality (measured in terms of iodine value and fatty acid composition). To analyze the map, they used a total of 252 markers (199 AFLP, 38 RFLP, and 15 SSR) with a map length of 1815 cM and with an average interval of 7 cM between adjacent markers, producing a total of 21 linkage groups with an average number of 12 markers per linkage groups. Again, almost in all maps, the markers were distributed at an interval of 25 cM except for linkage group 17 having 30 cM, indicating that the map is relatively homogeneous with regards to marker distribution; this is useful for tagging traits of economic interest for MAS. In this map, the length of individual linkage groups varied from 26.1 cM to 168 cM, with an average of 94 cM. The application of the genetic linkage map helps to detect QTLs for fatty acid composition in oil palm and serves as a tool for the MAS breeding program.

Similarly, a report from Seng [22] used a total of 120 hybrid crosses between high-yielding *dura* (ARK86D) x *pisifera* (ML161P) using AFLP markers. To construct the map, they used a total of 479 marker loci and 168 anchor points with a map length of 2247.5 cM and an average map density of 4.7 cM using LOD score of 3.0. They constructed a total of 16 linkage groups from 15-57 markers per linkage group with an average of 29 markers per linkage group and with the lengths ranging from 77.5 cM to 223.7 cM, and an average of 137 cM. In line with this, the markers were well distributed all over the 16 linkage groups. From their findings, Seng [22] concluded that the application of a genomic map in oil palm helps to validate against a closely related population and helps identify yield-related QTLs. Likewise, Ting *et al.* [30] and Ukoskit [23] also used the AFLP markers to construct a genetic linkage map in oil palm.

Simple sequence repeat (SSR) markers are co-dominant molecular markers that distinguish polymorphism and mapping in the oil palm genome. SSR markers were used for the first time to construct the map of the oil palm in the year 2005. To construct the map, Billotte *et al.* [17] used a total of 255 SSR markers with a map length of 1743 cM with an average marker density of 7 cM. using a LOD

score of 3.0 and producing a total of 16 linkage groups. Based on the outcome of their finding, mapping of oil palm using high-density makers like SSR brings comprehensive information for QTL mapping and other MAS research in the oil palm. In line with this, Billotte [20] used an SSR marker for QTL detection with a multi-parent linkage map of the cross between two oil palm populations. They used a total of 150 palms in the controlled cross between Africa (LM2T) x Deli (DA10D). To construct the map, a total of 251 SSR markers were used. Based on their finding, the SSR map for LM2T x DA10D had 16 linkage groups (LGs) and 253 loci, with a map length of 1479 cM and an average marker density of 6 cM. The large mapping genome was found in LG4 with spanned 134 cM on the average range of 61 - 250 cM and around 47% of the mapped loci having three or four alleles with an average density of 32 cM on the genome. In conclusion, a total of 156 SSRs (45%) and the *Sh* locus were mapped and the mapping of the crossed oil palm populations helped to identify the QTL for the major gene controlling fruit shell (*Sh*). By the same token, Montoya *et al.* [3] used a total of 347 segregating SSRs, 14 SNPs of genes, and the *Sh* locus to establish the linkage map and to detect QTLs of palm oil fatty acid composition. They produced a total of 16 LGs with a relative map length of 1485 cM and an average marker density of 4 cM at LOD 7.5 with a maximum recombination threshold of 0.3. Depending on their position on the linkage group, the length of the LGs ranged from 49.1 to 175.9 cM, with an average of 92.8 cM. Concerning QTLs, a total of 19 QTL associated with the palm oil fatty acid composition was obtained and this mapping helped to identify key genes in the oil palm genome related to oleic acid C18:1. In conclusion, 73% (253) of the mapped SSRs segregated only from the hybrid parent SA65T, 2%, (7) from PO3228D only, and 27% (93) were common SSRs segregating from both parents. Again, the high number of mapped SSR loci with accurate relative linear orders, and their molecular hyper-variability helped to undertake other such mapping studies in other *Elaeis* breeding materials. Later on, Cochard [21] constructed a linkage map using a 281 SSRs marker and a total of 271 genotyped oil palm populations. They produced a total of 16 linkage groups covering group A (2078 cM) and group B (1845 cM), with an average density of one marker every 9 and 7 cM, respectively. Generally, the integrated maps gave a total map length of 1935 cM with a total of 281 markers and an average density of one marker for every 7.4 cM. Besides, the marker orders between physical and genetic maps were in good accordance, except for some sporadic markers. Based upon their finding they concluded that this output could help to step towards efficient pedigree-based quantitative trait locus (QTL) mapping using the first intercrossed generations in current breeding programs. Similar studies have been done using SSRs for the mapping of the oil palm genome. For instance, QTLs identification is associated with callogenesis and embryogenesis [30], QTL mapping for oil yield using African oil palm [54], linkage map and QTL analysis for sex ratio and related traits [23], genetic map construction for two independent oil palm hybrids [55], linkage mapping and identifica-

tion of major QTL genes for stem height [15] which all brought remarkable information for the oil palm genome mapping and molecular breeding research.

Currently, oil palm single nucleotide polymorphisms (SNPs) are the most highly preferred and high-density markers used to study genetic diversity and population structure, to construct high-density genetic maps, and provide genotypes for the genome-wide association [56], and genomic selection studies [5]. The first SNP marker-based oil palm genome mapping was constructed by Jeennor and Volkaert [54] using a total of 190 segregating loci (89 SSRs, 90 genes, and 11 non-gene based SNP markers), which were mapped into 31 linkage groups by applying threshold LOD of 3 and a recombination fraction of 0.45. They produced a map with a total length of 1233 cM containing two to 20 markers covering a length between 1.5 and 103.5 cM, and with an average distance between markers of 6.5 cM. This finding helped to identify validated candidate genes involved in lipid biosynthesis and mapped near significant QTL for various economic yield traits. This indicates the applicability of markers for MAS to improve the required trait selection for the oil palm breeding programs.

Moreover, Pootakham [27] developed SNP markers using the GBS method in the African oil palm with a total of 1085 SNPs to construct a linkage map. The map produced spanned 1429.6 cM and had an average of one marker every 1.26 cM. They also detected on LG 10, 14, and 15, three QTL genes affecting trunk height whereas a single QTL associated with fruit bunch weight was identified on LG 3. They concluded that mapping of oil palm genome by the use of Genotyping by sequencing (GBS) approach helped to produce high-density maps and could enhance knowledge on genome structure which is valuable for mapping other economically important genes for MAS. Bai *et al.* [28] also used high-density GBS marker data to construct and detect QTL associated with leaf area using 145 oil palm breeding populations derived from a cross between Deli Dura and Avros Pisifera. They constructed a genetic linkage map using a total of 2413 SNPs, producing a total of 16 linkage groups with a total length of 1161.89 cM, and an average marker spacing of 0.48 cM. Based on their results, two potential QTL for leaf area were detected on Chr 3 and 9 and the gene ARC5, located in the QTL region on Chr 9, was the most likely candidate gene responsible for leaf growth in oil palm. They concluded that the use of a high-quality and SNP-based map supplies a base to fine map QTL for agronomic traits and MAS yield improvement in oil palm.

Furthermore, Gan [24] reported the first DArT-based genetic linkage maps using two closely related oil palm populations. For this purpose, they used a total of 1399 DArT and 1466 SNP markers. They produced a total of 16 major independent linkage groups with map lengths of 1873.7 and 1720.6 cM and with an average marker density of 1.34 and 1.17 cM, respectively. The integrated map was 1803.1 cM long with 2066 mapped markers and an average marker density of 0.87 cM. In conclusion, the use of high-density marker DArTseq marker

helped to generate high-density genetic maps in oil palm, and the integration of maps was also useful to study QTL analysis of important yield traits and other MAS studies. By the same token, Ong [25] also reported a linkage-based genome assembly in oil palm. To construct the map, they used a total of 27,890 SNP markers and generated a total of 16 linkage groups with a total map length of 1151.7 cM and an average mapping interval of 0.04 cM. This mapping helped to study QTLs in sugar and lipid biosynthesis pathways. It also helped to improve knowledge on the current physical genome of the commercial oil palm.

Recently, Single Primer Enrichment Technology (SPET) markers were used to construct a high-density genetic linkage map from a controlled cross of two oil palm (*Elaeis guineensis*) genotypes [12]. To construct the map, they used a total of 3501 SPET markers with a total length of 1370 cM and 1.74 markers per cM (0.57 cM/marker). This resulted in a total of 16 linkage groups with a total of 1054 loci. From their work, they concluded that the application of these cost-efficient SPET markers are suitable for linkage map construction in oil palm and probably, also in other species.

6. Factors Limiting Oil Palm Genome Mapping

For the last two decades, numerous findings in the area of oil palm genome mapping have been reported by different scholars (Table 3). These have remarkably improved knowledge on the genetic improvement of the oil palm using marker-assisted selection strategies. However, the success of genome mapping of the oil palm is highly dependent on several factors. For instance, Mayes *et al.* [14] reported that the choice of mapping populations is one of the major determinant factors of genome mapping of the oil palm, stating that to select them, several criteria were to be considered such as the simplicity of cross for allele scoring and linkage analysis, representation of alleles within breeding materials, and availability of phenotypic data. The variation is very clear between Asian and African types of oil palm genetic materials. Again, another factor affecting genome mapping is the genetic marker type used for mapping. In this regard, a report (Table 3), clearly showed that marker polymorphism creates a variation in the outcome of genetic linkage groups. For instance, the first RFLP marker by Mayes *et al.* [14] produced a total of 24 LGs, while the first SSR marker by Billotte *et al.* [17] produced 16 LGs. Very recently, the use of high-density markers like SNPs brought more light to the genetic mapping of oil palm. In addition to the type of marker, the density of markers used also brought variation in genome mapping of the oil palm. For example, Mayes *et al.* [14] used a total of 97 RFLP markers while Rance [29] used a total of 153 RFLP markers. Moreover, a total of 49 additional marker loci resulted in an improvement of map resolution from 24 linkage groups [14] to 22 linkage groups [29] and not only the map resolution but, also the total map length differed; as the number of markers increased from 97 to 153, the map length decreased from 860 cM to 852 cM, respectively. Conversely, Billotte *et al.* [17] used a combination of 255 SSR and 688

AFLP markers, based on this and due to the high-density markers used (AFLPs); 23% of the filled gaps were covered by this marker relative to the SSR based map. In the same vein, compared to the results of Billotte *et al.* [17] *i.e.*, 255 SSRs in combination with other low-density markers resulting in a total length of 1743 cM with an average marker density of 7 cM, Billotte [20] recently used independent high-density markers *i.e.*, 251 SSR and obtained a total map length of 1479 cM with an average marker density of 6 cM. The variation is clear that in the later, they used a single high-density mapping marker.

Besides, the population sample size is another factor that brings a variation in the genome mapping of the oil palm. Singh [18] reported that even though they are using high numbers of markers, this doesn't result in a fine map. Based on their conclusion, this result is due to the use of a small sample size of the F1 progeny. Equally, Billotte [20] reported that there is a variation of map detection power, due to the variation in population size, and based on their report, a large population size of the multi-parent system provides greater detection power for the QTL than biparental and small populations. By the same token, Ukoskit [23] also reported that the difference in map length is due to the variation in the population size. In general, a large number of markers with large population sizes (pedigree populations) results in better genome mapping [57].

Last but not the least, genome mapping is also highly governed by the software used for mapping the genome. In the oil palm, various software programs have been used to build genetic linkage maps (Table 3), such as MAPMAKER 2.0 [58], JoinMap ver 2.0 [59], CRI-MAP [58], JoinMap ver 3.0 [60], JoinMap ver 4.0 [61], Lep-MAP 2 [62] and LepMAP 3 [57]. Due to the perennial nature of the oil palm, its out-crossing nature and long generation time result in difficulty to obtain enough genetic materials or mapping populations, to overcome these limitations, consensus genetic maps are obtained by integrating multiple unrelated genetic maps sharing common markers. Such consensus maps can be constructed by different linkage map software (Table 3). Nowadays, Lep-MAP3 (LM3) is a novel linkage map construction software suite. It can handle millions of markers and thousands of individuals possibly from multiple families [57].

7. Conclusion

In perennial crops like oil palm, getting new or improved varieties through conventional breeding methods is difficult because it is time-consuming and costly, all related to the long generation cycles, large plant size, and the long evaluation period of 10 - 15 years. The application of marker-assisted breeding techniques in this crop helps to minimize the above-listed constraints. The construction of genetic linkage maps plays a major role in the genetic analysis and molecular breeding programs of oil palm. This has been used for the identification of genetic loci using different traits such as yield and its components, oil quality, and abiotic stress, resulting in better genetic improvement and more cost-effective breeding. Nowadays, high-throughput molecular markers sequencing technolo-

gy helps to raise both genetic and physical maps to a new level by providing an increased sequence pool from which to build genetic maps and assemble genome sequences. In the last two decades, most of the research findings reveal that genome mapping helps in the identification of major genes that control quantitative traits like yield and quality of oil palm. Furthermore, the literature on this crop shows that there is a variation of genome mapping due to several factors; for instance, marker type, marker density, size of the mapped population, and software used. Despite all pros and cons, genome mapping in this crop plays a crucial role, and to get a more pronounced map in the future, oil palm genome mapping should focus on the use of high-density molecular marker types, a large number of mapping population, and up-to-date software that can yield remarkable results and help to map and detect more quantitative traits related to both yield and oil quality.

Authors' Contributions

Essubalew Getachew SEYUM: Developed an idea and wrote the manuscript whereas all others are involved by commenting, suggesting, and re-arranging the setup of the manuscript.

Acknowledgements

The authors acknowledge the GENES program of the Intra-Africa Academic Mobility Scheme of the European Union for financial support (EU-GENES: 2017-2552/001-001).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

- [1] Cros, D., Denis, M., Bouvet, J.M. and Sánchez, L. (2015) Long-Term Genomic Selection for Heterosis without Dominance in Multiplicative Traits: Case Study of Bunch Production in Oil Palm. *BMC Genomics*, **16**, Article No. 651. <https://doi.org/10.1186/s12864-015-1866-9>
- [2] Hartley, C.W.S. (1988) *The Oil Palm (Elaeis guineensis Jacq.)*. Longman Scientific & Technical, New York.
- [3] Montoya, C., Lopes, R., Flori, A., Cros, D., Cuellar, T., Summo, M., Espeout, S., Rivallan, R., Risterucci, A.M., Bittencourt, D., Zambrano, J.R., Alarcón, G.W.H., Villeneuve, P., Pina, M., Nouy, B., Amblard, P., Ritter, E., Leroy, T. and Billotte, N. (2013) Quantitative Trait Loci (QTLs) Analysis of Palm Oil Fatty Acid Composition in an Interspecific Pseudo-Backcross from *Elaeis oleifera* (H.B.K.) Cortés and Oil Palm (*Elaeis guineensis* Jacq.). *Tree Genetics and Genomes*, **9**, 1207-1225. <https://doi.org/10.1007/s11295-013-0629-5>
- [4] Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., Cui, Z., Ruan, Y., Burgueño, J., San Vicente, F., Olsen, M., Prasanna, B.M., Crossa, J., Yu, H. and Zhang, X. (2017) Effect of Trait Heritability, Training Population Size and Marker

- Density on Genomic Prediction Accuracy Estimation in 22 Bi-Parental Tropical Maize Populations. *Frontiers in Plant Science*, **8**, Article No. 1916. <https://doi.org/10.3389/fpls.2017.01916>
- [5] Cros, D., Tchounke, B. and Nkague-Nkamba, L. (2018) Training Genomic Selection Models across Several Breeding Cycles Increases Genetic Gain in Oil Palm *in Silico* Study. *Molecular Breeding*, **38**, Article No. 89. <https://doi.org/10.1007/s11032-018-0850-x>
- [6] Wong, C.K. and Bernardo, R. (2008) Genomewide Selection in Oil Palm: Increasing Selection Gain per Unit Time and Cost with Small Populations. *Theoretical and Applied Genetics*, **116**, 815-824. <https://doi.org/10.1007/s00122-008-0715-5>
- [7] USDA (2020). <https://www.fas.usda.gov/data/oilseeds-world-markets-and-trade>
- [8] Babu, B.K. and Mathur, R.K. (2016) Molecular Breeding in Oil Palm (*Elaeis guineensis*): Status and Future Perspectives. *Progressive Horticulture*, **48**, 123-131. <https://doi.org/10.5958/2249-5258.2016.00051.8>
- [9] Corley, R.H.V. and Tinker, P.B. (2016) The Oil Palm. Wiley-Blackwell, Hoboken. <https://doi.org/10.1002/9781118953297>
- [10] Soh, A.C., Mayes, S. and Roberts, J.A. (2017) Oil Palm Breeding: Genetics and Genomics. 1st Edition, CRC Press, Boca Raton. <https://doi.org/10.1201/9781315119724-1>
- [11] 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. and Appleton, D.R. (2016) Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm. *Molecular Plant*, **9**, 1132-1141. <https://doi.org/10.1016/j.molp.2016.04.010>
- [12] Herrero, J., Santika, B., Herrán, A., Erika, P., Sarimana, U., Wendra, F., Sembiring, Z., Asmono, D. and Ritter, E. (2020) Construction of a High Density Linkage Map in Oil Palm Using SPET Markers. *Scientific Reports*, **10**, Article No. 9998. <https://doi.org/10.1038/s41598-020-67118-y>
- [13] Xu, Y. and Crouch, J.H. (2008) Marker-Assisted Selection in Plant Breeding: From Publications to Practice. *Crop Science*, **48**, 391-407. <https://doi.org/10.2135/cropsci2007.04.0191>
- [14] Mayes, S., Jack, P.L., Corley, R.H. and Marshall, D.F. (1997) Construction of a RFLP Genetic Linkage Map for Oil Palm (*Elaeis guineensis* Jacq.). *Genome*, **40**, 116-122. <https://doi.org/10.1139/g97-016>
- [15] Lee, M., Xia, J.H., Zou, Z., Ye, J., Rahmadsyah, Alfiko, Y., Jin, J., Lieando, J.V., Purnamasari, M.I., Lim, C.H., Suwanto, A., Wong, L., Chua, N.H. and Yue, G.H. (2015) A Consensus Linkage Map of Oil Palm and a Major QTL for Stem Height. *Scientific Reports*, **5**, Article No. 8232. <https://doi.org/10.1038/srep08232>
- [16] O'Rourke, J.A. (2014) Genetic and Physical Map Correlation. *eLS*, Wiley-Blackwell, Hoboken, 1-4. <https://doi.org/10.1002/9780470015902.a0000819.pub3>
- [17] Billotte, N., Marseillac, N., Risterucci, A.M., Adon, B., Brottier, P., Baurens, F.C., Singh, R., Herrán, A., Asmady, H., Billot, C., Amblard, P., Durand-Gasselín, T., Courtois, B., Asmono, D., Cheah, S.C., Rohde, W., Ritter, E. and Charrier, A. (2005) Microsatellite-Based High Density Linkage Map in Oil Palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics*, **110**, 754-765. <https://doi.org/10.1007/s00122-004-1901-8>
- [18] Singh, R., Tan, S.G., Panandam, J.M., Rahman, R.A., Ooi, L.C., Low, E.T.L., Sharma, M., Jansen, J. and Cheah, S.C. (2009) Mapping Quantitative Trait Loci (QTLs) for

- Fatty Acid Composition in an Interspecific Cross of Oil Palm. *BMC Plant Biology*, **9**, Article No. 114. <https://doi.org/10.1186/1471-2229-9-114>
- [19] Moretzsohn, M.C., Nunes, C.D.M., Ferreira, M.E. and Grattapaglia, D. (2000) RAPD Linkage Mapping of the Shell Thickness Locus in Oil Palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics*, **100**, 63-70. <https://doi.org/10.1007/s001220050009>
- [20] Billotte, N., Jourjon, M.F., Marseillac, N., Berger, A., Flori, A., Asmady, H., Adon, B., Singh, R., Nouy, B., Potier, F., Cheah, S.C., Rohde, W., Ritter, E., Courtois, B., Charrier, A. and Mangin, B. (2010) QTL Detection by Multi-Parent Linkage Mapping in Oil Palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics*, **120**, 1673-1687. <https://doi.org/10.1007/s00122-010-1284-y>
- [21] Cochard, B., Carrasco-Lacombe, C., Pomies, V., Dufayard, J.F., Suryana, E., Omoré, A., Tristan, D.G. and Tisné, S. (2015) Pedigree-Based Linkage Map in Two Genetic Groups of Oil Palm. *Tree Genetics and Genomes*, **11**, Article No. 68. <https://doi.org/10.1007/s11295-015-0893-7>
- [22] Seng, T.Y., Saad, S.H.M., Chin, C.W., Ting, N.C., Singh, R.S.H., Zaman, F.Q., Tan, S.G. and Alwee, S.S.R.S. (2011) Genetic Linkage Map of a High Yielding FELDA Deli×Yangambi Oil Palm Cross. *PLoS ONE*, **6**, e26593. <https://doi.org/10.1371/journal.pone.0026593>
- [23] Ukoskit, K., Chanroj, V., Bhusudsawang, G., Pipatchartlearnwong, K., Tangphatsornruang, S. and Tragoonrung, S. (2014) Oil Palm (*Elaeis guineensis* Jacq.) Linkage Map, and Quantitative Trait Locus Analysis for Sex Ratio and Related Traits. *Molecular Breeding*, **33**, 415-424. <https://doi.org/10.1007/s11032-013-9959-0>
- [24] Gan, S.T., Wong, W.C., Wong, C.K., Soh, A.C., Kilian, A., Low, E.T.L., Massawe, F. and Mayes, S. (2018) High Density SNP and DArT-Based Genetic Linkage Maps of Two Closely Related Oil Palm Populations. *Journal of Applied Genetics*, **59**, 23-34. <https://doi.org/10.1007/s13353-017-0420-7>
- [25] Ong, A.L., Teh, C.K., Kwong, Q.B., Tangaya, P., Appleton, D.R., Massawe, F. and Mayes, S. (2019) Linkage-Based Genome Assembly Improvement of oil Palm (*Elaeis guineensis*). *Scientific Reports*, **9**, Article No. 6619. <https://doi.org/10.1038/s41598-019-42989-y>
- [26] Ong, A.L., Teh, C.K., Mayes, S., Massawe, F., Appleton, D.R. and Kulaveerasingam, H. (2020) An Improved Oil Palm Genome Assembly as a Valuable Resource for Crop Improvement and Comparative Genomics in the *Arecoideae* Subfamily. *Plants*, **9**, Article No. 1476. <https://doi.org/10.3390/plants9111476>
- [27] Pootakham, W., Jomchai, N., Ruang-areerate, P., Shearman, J.R., Sonthirod, C., Sangsrakru, D., Tragoonrung, S. and 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**, 288-295. <https://doi.org/10.1016/j.ygeno.2015.02.002>
- [28] Bai, B., Zhang, Y.J., Wang, L., Lee, M., Rahmadsyah, Ye, B.Q., Alfiko, Y., Purwantomo, S., Suwanto, A. and Yue, G.H. (2018) Mapping QTL for Leaf Area in Oil Palm Using Genotyping by Sequencing. *Tree Genetics and Genomes*, **14**, Article No. 31. <https://doi.org/10.1007/s11295-018-1245-1>
- [29] Rance, K.A., Mayes, S., Price, Z., Jack, P.L. and Corley, R.H.V. (2001) Quantitative Trait Loci for Yield Components in Oil Palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics*, **103**, 1302-1310. <https://doi.org/10.1007/s122-001-8204-z>
- [30] Ting, N.C., Jansen, J., Nagappan, J., Ishak, Z., Chin, C.W., Tan, S.G., Cheah, S.C. and Singh, R. (2013) Identification of QTLs Associated with Callogenesis and Em-

- bryogenesis in Oil Palm Using Genetic Linkage Maps Improved with SSR Markers. *PLoS ONE*, **8**, e53076. <https://doi.org/10.1371/journal.pone.0053076>
- [31] Meksem, K., Ishihara, H. and Jesse, T. (2005) Integration of Physical and Genetic Maps. In: Meksem, K. and Kahl, G., Eds., *The Handbook of Plant Genome Mapping: Genetic and Physical Mapping*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 215-232. <https://doi.org/10.1002/3527603514.ch9>
- [32] Meksem, K. and Kahl, G. (2005) *The Handbook of Plant Genome Mapping: Genetic and Physical Mapping*. 1st Edition, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim. <https://doi.org/10.1002/3527603514>
- [33] Dixit, R., Jayanand, D., Rai, D., Agarwal, R. and Pundhir, A. (2014) Physical Mapping of Genome and Genes. *Journal of Biological Engineering Research and Review*, **1**, 6-11.
- [34] Shah, M., Varshney, P., Patel, P., Patel, D. and Meshram, D. (2012) Cytogenetic Mapping Techniques: An Approach to Genome Analysis. *Research & Reviews in BioSciences*, **7**, 209-219.
- [35] Hozier, J.C. and Davis, L.M. (1992) Cytogenetic Approaches to Genome Mapping. *Analytical Biochemistry*, **200**, 205-217. [https://doi.org/10.1016/0003-2697\(92\)90455-G](https://doi.org/10.1016/0003-2697(92)90455-G)
- [36] Azhaguvel, P., Weng, Y., Babu, R., Manickavelu, A., Saraswathi, D. and Balyan, H. (2010) Fundamentals of Physical Mapping. In: Kole, C. and Abbott, A.G., Eds., *Principles and Practices of Plant Genomics*, CRC Press, Boca Raton, 24-62.
- [37] Brown, T.A. (2002) *Mapping Genomes*, Genomes. 2nd Edition, Wiley-Liss, Hoboken.
- [38] Deonier, R.C., Waterman, M.S. and Tavaré, S. (2005) Physical Mapping of DNA. *Computational Genome Analysis: An Introduction*. Springer, New York, 99-119. https://doi.org/10.1007/0-387-28807-4_4
- [39] Hass-Jacobus, B. and Jackson, S.A. (2005) Physical Mapping of Plant Chromosomes. In: Meksem, K. and Kahl, G., Eds., *The Handbook of Plant Genome Mapping*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 131-149. <https://doi.org/10.1002/3527603514.ch6>
- [40] Lewin, H.A., Larkin, D.M., Pontius, J. and O'Brien, S.J. (2009) Every Genome Sequence Needs a Good Map. *Genome Research*, **19**, 1925-1928. <https://doi.org/10.1101/gr.094557.109>
- [41] Griffiths, A.J., Miller, J.H., Suzuki, D.T., Lewontin, R.C. and Gelbart, W.M. (2000) *An Introduction to Genetic Analysis*. 7th Edition, W.H. Freeman, New York.
- [42] Li, Y. (2015) Construction of a High-Density High-Resolution Genetic Map and Its Integration with BAC-Based Physical Map in Channel Catfish. *DNA Research*, **22**, 39-52. <https://doi.org/10.1093/dnares/dsu038>
- [43] Sturtevant, A.H. (1913) The Linear Arrangement of Six Sex-Linked Factors in *Drosophila*, as Shown by Their Mode of Association. *Journal of Experimental Zoology*, **14**, 43-59. <https://doi.org/10.1002/jez.1400140104>
- [44] Alves, J.M., Chikhi, L., Amorim, A. and Lopes, A.M. (2014) The 8p23 Inversion Polymorphism Determines Local Recombination Heterogeneity across Human Populations. *Genome Biology and Evolution*, **6**, 921-930. <https://doi.org/10.1093/gbe/evu064>
- [45] Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980) Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *American Journal of Human Genetics*, **32**, 314-331.
- [46] Williams, J.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S.V. (1990) DNA

- Polymorphisms Amplified by Arbitrary Primers Are Useful as Genetic Markers. *Nucleic Acids Research*, **18**, 6531-6535. <https://doi.org/10.1093/nar/18.22.6531>
- [47] Litt, M. and Luty, J.A. (1989) A Hypervariable Microsatellite Revealed by *in Vitro* Amplification of a Dinucleotide Repeat within the Cardiac Muscle Actin Gene. *American Journal of Human Genetics*, **44**, 397-401.
- [48] Palazzolo, M.J., Sawyer, S.A., Martin, C.H., Smoller, D.A. and Hartl, D.L. (1991) Optimized Strategies for Sequence-Tagged-Site Selection in Genome Mapping. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 8034-8038. <https://doi.org/10.1073/pnas.88.18.8034>
- [49] Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J. and Kuiper, M. (1995) AFLP: A New Technique for DNA Fingerprinting. *Nucleic Acids Research*, **23**, 4407-4414. <https://doi.org/10.1093/nar/23.21.4407>
- [50] Lai, E., Riley, J., Purvis, I. and Roses, A. (1998) A 4-Mb High-Density Single Nucleotide Polymorphism-Based Map around Human APOE. *Genomics*, **54**, 31-38. <https://doi.org/10.1006/geno.1998.5581>
- [51] Williams, M.N.V., Pande, N., Nair, S., Mohan, M. and Bennett, J. (1991) Restriction Fragment Length Polymorphism Analysis of Polymerase Chain Reaction Products Amplified from Mapped Loci of Rice (*Oryza sativa* L.) Genomic DNA. *Theoretical and Applied Genetics*, **82**, 489-498. <https://doi.org/10.1007/BF00588604>
- [52] Lyamichev, V., Brow, M.A. and Dahlberg, J.E. (1993) Structure-Specific Endonucleolytic Cleavage of Nucleic Acids by Eubacterial DNA Polymerases. *Science*, **260**, 778-783. <https://doi.org/10.1126/science.7683443>
- [53] Bai, B., Wang, L., Zhang, Y.J., Lee, M., Rahmadsyah, R., Alfiko, Y., Ye, B.Q., Purwantomo, S., Suwanto, A., Chua, N.H. and Yue, G.H. (2018) Developing Genome-Wide SNPs and Constructing an Ultrahigh-Density Linkage Map in Oil Palm. *Scientific Reports*, **8**, Article No. 691. <https://doi.org/10.1038/s41598-017-18613-2>
- [54] Jeennor, S. and Volkaert, H. (2014) Mapping of Quantitative Trait Loci (QTLs) for Oil Yield Using SSRs and Gene-Based Markers in African Oil Palm (*Elaeis guineensis* Jacq.). *Tree Genetics and Genomes*, **10**, 1-14. <https://doi.org/10.1007/s11295-013-0655-3>
- [55] Ting, N.C., Jansen, J., Mayes, S., Massawe, F., Sambanthamurthi, R., Ooi, L.C.L., Chin, C.W., Arulandoo, X., Seng, T.Y., Alwee, S.S.R.S., Ithnin, M. and Singh, R. (2014) High Density SNP and SSR-Based Genetic Maps of Two Independent Oil Palm Hybrids. *BMC Genomics*, **15**, Article No. 309. <https://doi.org/10.1186/1471-2164-15-309>
- [56] Xia, W., Luo, T., Zhang, W., Mason, A.S., Huang, D., Huang, X., Tang, W., Dou, Y., Zhang, C. and Xiao, Y. (2019) Development of High-Density SNP Markers and Their Application in Evaluating Genetic Diversity and Population Structure in *Elaeis guineensis*. *Frontiers in Plant Science*, **10**, Article No. 130. <https://doi.org/10.3389/fpls.2019.00130>
- [57] Rastas, P. (2017) Lep-MAP3: Robust Linkage Mapping Even for Low-Coverage Whole Genome Sequencing Data. *Bioinformatics*, **33**, 3726-3732. <https://doi.org/10.1093/bioinformatics/btx494>
- [58] Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E. and Newburg, L. (1987) MAPMAKER: An Interactive Computer Package for Constructing Primary Genetic Linkage Maps of Experimental and Natural Populations. *Genomics*, **1**, 174-181.
- [59] Stam, P. (1993) Construction of Integrated Genetic Linkage Maps by Means of a

New Computer Package: Join Map. *The Plant Journal*, **3**, 739-744.

- [60] Ooijen, J.W. and Voorrips, R.E. (2002) JoinMap: Version 3.0: Software for the Calculation of Genetic Linkage Maps. University and Research Center.
- [61] Van Ooijen, J.W. (2006) JoinMap® 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations. Kyazma B.V., Wagening.
- [62] Rastas, P., Paulin, L., Hanski, I., Lehtonen, R. and Auvinen, P. (2013) Lep-MAP: Fast and Accurate Linkage Map Construction for Large SNP Datasets. *Bioinformatics*, **29**, 3128-3134. <https://doi.org/10.1093/bioinformatics/btt563>



Genome properties of key oil palm (*Elaeis guineensis* Jacq.) breeding populations

Essubalew Getachew Seyum^{1,2,3} · Ngalle Hermine Bille¹ · Wosene Gebreselassie Abteaw³ · Pasi Rastas⁴ · Deni Arifianto⁵ · Hubert Domonhédou⁶ · Benoît Cochard⁷ · Florence Jacob⁷ · Virginie Riou^{8,9} · Virginie Pomiès^{8,9} · David Lopez^{8,9} · Joseph Martin Bell¹ · David Cros^{8,9}

Received: 28 January 2022 / Revised: 26 May 2022 / Accepted: 4 June 2022
© The Author(s), under exclusive licence to Institute of Plant Genetics Polish Academy of Sciences 2022

Abstract

A good knowledge of the genome properties of the populations makes it possible to optimize breeding methods, in particular genomic selection (GS). In oil palm (*Elaeis guineensis* Jacq), the world's main source of vegetable oil, this would provide insight into the promising GS results obtained so far. The present study considered two complex breeding populations, Deli and La Mé, with 943 individuals and 7324 single-nucleotide polymorphisms (SNPs) from genotyping-by-sequencing. Linkage disequilibrium (LD), haplotype sharing, effective size (N_e), and fixation index (F_{st}) were investigated. A genetic linkage map spanning 1778.52 cM and with a recombination rate of 2.85 cM/Mbp was constructed. The LD at $r^2=0.3$, considered the minimum to get reliable GS results, spanned over 1.05 cM/0.22 Mbp in Deli and 0.9 cM/0.21 Mbp in La Mé. The significant degree of differentiation existing between Deli and La Mé was confirmed by the high F_{st} value (0.53), the pattern of correlation of SNP heterozygosity and allele frequency among populations, and the decrease of persistence of LD and of haplotype sharing among populations with increasing SNP distance. However, the level of resemblance between the two populations over short genomic distances (correlation of r values between populations >0.6 for SNPs separated by <0.5 cM/1 kbp and percentage of common haplotypes $>40\%$ for haplotypes <3600 bp/0.20 cM) likely explains the superiority of GS models ignoring the parental origin of marker alleles over models taking this information into account. The two populations had low N_e (<5). Population-specific genetic maps and reference genomes are recommended for future studies.

Keywords Genome properties · Genomic selection · Genotyping-by-sequencing · Oil palm · Single nucleotide polymorphisms

Communicated by: Izabela Pawłowicz

✉ David Cros
david.cros@cirad.fr

¹ Department of Plant Biology and Physiology, Faculty of Sciences, University of Yaoundé I, Yaoundé, Cameroon

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

³ Department of Horticulture and Plant Sciences, Jimma University College of Agriculture and Veterinary Medicine, P.O. Box 307, Jimma, Ethiopia

⁴ Institute of Biotechnology, Helsinki Institute of Life Science (HiLIFE), University of Helsinki, 00014 Helsinki, Finland

⁵ P.T. SOCFINDO Medan, Medan, Indonesia

⁶ INRAB, CRA-PP, Pobè, Benin

⁷ PalmElit SAS, Montferrier sur Lez, France

⁸ CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), UMR AGAP Institut, F-34398 Montpellier, France

⁹ UMR AGAP Institut, University of Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France

Introduction

The African oil palm (*Elaeis guineensis* Jacq.) is a perennial tropical monocot oil-producing plant that belongs to the *Arecaceae* family. It originated from the Gulf of Guinea. It is naturally cross-pollinated, monoecious, allogamous, and diploid, with a chromosome number of $2n = 2x = 32$ and having a genome sequence of 1.8 gigabases (Ithnin and Din 2020). The economic life span of oil palm ranges from 25 to 30 years and it is mainly cultivated in humid tropical zones of the world (Barcelos et al. 2015).

The total world vegetable oil production is currently around 200 million metric tons (MT), led by oil palm (37.5%), followed by soybean oil (30%), rapeseed oil (14%), and sunflower oil (9.5%) (Statista 2021). The world demand for oil palm is expected to reach 240 million tons by 2050 (Corley 2009). Oil palm produces an average oil yield of 4 tons per hectare every year, which is approximately 7–10 times higher than soybean (Babu and Mathur 2016; Corley and Tinker 2016; Pirker et al. 2016). Oil palm is an important source of edible oil with over 80% of the products used in the food industry (cooking/frying oil, shortenings, margarine, and confectionery fats), and the rest used in the chemical industry for the formulation of soaps and detergents, pharmaceutical products, cosmetics, biodiesel, etc. (Basiron 2007; Corley 2009; Soh et al. 2017).

Nowadays, the cultivation of oil palm relies on hybrid varieties because they have a high yield per hectare. Group A and group B are the two heterotic groups involved in the development of hybrid cultivars of African oil palm (Nyouma et al. 2019). Group A mostly consists of the Deli parental population, which is derived from four individuals from an unknown area of Africa planted in 1848 in Indonesia (Hartley 1988). The selection of the Deli population, mainly for yield, started in the early twentieth century. Group B is made up of several African breeding populations. African populations resulted from a limited number of founders collected during the first half of the twentieth century. La Mé population originated from individuals collected in the Bingerville region of the Ivory Coast between 1924 and 1930, with three founders for the La Mé individuals considered here (Cochard et al. 2009). In both A and B groups, inbreeding was commonly used, by using selfing or by mating with related selected individuals (Corley and Tinker 2016).

Despite its wide adaptation and importance, oil palm production and productivity are generally far from their potential due to biotic and abiotic practical constraints. Climate change, land shortage, labor shortage, and diseases (in particular vascular wilt, ganoderma, and bud rot) are among the major factors hindering the current and

future yield of oil palm across the world (Corley 2009; Paterson et al. 2013; Barcelos et al. 2015; Kwong et al. 2016; Pirker et al. 2016). In addition, genetic improvement through the conventional method in oil palm is constrained by several factors, in particular a long breeding cycle (>15 years) and a limited number of tested individuals (Cros et al. 2015; Jin et al. 2016; Seng et al. 2016). To provide a solution while ensuring a sustainable future, marker-assisted breeding has recently been introduced into oil palm breeding programs (Soh et al. 2017) with, for instance, the detection of quantitative trait loci (QTLs) controlling oil yield, quality, vegetative growth, and resistance to diseases (Pootakham et al. 2015; Tisé et al. 2015; Ithnin et al. 2017; Bai et al. 2018b; Daval et al. 2021) and genomic selection (Nyouma et al. 2019).

Genomic selection (GS) is a marker-assisted selection (MAS) method with a high density of markers on the entire genome so that at least one marker is in linkage disequilibrium with each QTL (Meuwissen et al. 2001). It is the most effective MAS method to improve quantitative traits (Heffner et al. 2009). Studies on the application of GS in oil palm brought positive results. Thus, GS could improve oil palm clonal selection (Nyouma et al. 2020) and the selection of parents to use for hybrid crossings (Cros et al. 2017). Generally, GS in oil palm can enhance selection intensity and/or shorten the generation interval, thus increasing the annual genetic gain (Nyouma et al. 2019). The different studies carried out have provided a significant amount of information concerning the conditions of implementation of GS in this species. For example, in Deli and La Mé, GS has been implemented with relatively small training populations (<150) and low marker density (<2000 SNPs) (Cros et al. 2017; Nyouma et al. 2020); and models ignoring the parental origin of marker alleles were found to be more accurate than models accounting for this information (Nyouma et al. 2020, 2022). To better understand the GS results in the populations involved, a detailed study of their genome properties would be of interest, particularly regarding linkage disequilibrium (LD), effective size (N_e), haplotype sharing, and fixation index (F_{st}), which are known to affect GS accuracy.

Linkage disequilibrium is defined as the non-random association of alleles at two or more loci (Weir 1979; Slatkin 2008). The concept of GS relies heavily on LD between QTLs and DNA markers, and a good knowledge of LD in the breeding population is necessary to optimize GS (Nakaya and Isobe 2012; Technow et al. 2014; Li and Kim 2015; Bejarano et al. 2018). The LD pattern is shaped by genetic factors, i.e., mutations and historical events that occurred during domestication and population formation, including natural and artificial selection, drift, migration, and non-random mating, as well as by non-genetic factors such as marker ascertainment bias (Flint-Garcia et al. 2003; Gupta et al. 2005; Mackay and Powell 2007; Slatkin 2008).

Effective size (N_e) is defined as the size of an idealized Wright-Fisher population that would give rise to the same extent of random genetic drift or rate of inbreeding as the actual population (Wright 1931; Caballero 1994; Falconer and Mackay 1996). Low N_e leads to high rates of genetic drift and inbreeding in a population, making N_e one of the major factors influencing LD, and consequently the accuracy of GS (Grattapaglia 2014). N_e can be inferred from LD (Corbin et al. 2012), as there is an inverse relationship between LD and N_e . For a given marker density, training population size, and trait, LD and GS prediction accuracy are higher in populations with low N_e than in populations with high N_e (Solberg et al. 2008; Daetwyler et al. 2010; Wientjes et al. 2013; Grattapaglia 2014; Lin et al. 2014). So far, in oil palm, N_e was only estimated in the Deli population (Cros et al. 2014) and there is no information about N_e for La Mé population.

Haplotypes correspond to two or more SNP alleles that tend to be inherited as a unit in the chromosome (Bernardo 2010). Haplotype sharing helps estimate the genetic resemblance between individuals and is a natural extension of identity by descent (Xu and Guan 2014). Several authors showed that the aggregation of SNPs into haplotypes can increase the prediction accuracy in animals (Calus et al. 2008; Cuyabano et al. 2014; Teissier et al. 2020) and in plant species that were allogamous or with high multiallelism (Matias et al. 2017; Ballesta et al. 2019). Also, consistency of linkage phases between QTL and marker alleles among populations is required to pool them to get a larger population for genetic studies (De Roos et al. 2009; Technow et al. 2012).

The fixation index (F_{st}) is the correlation between gametes chosen randomly from within the same sub-population relative to the entire population (Wright 1931; Jakobsson et al. 2013; Weir and Goudet 2017). It is used to identify loci with divergent allelic frequencies between two or more populations. It helps to understand the genetic differentiation among groups. It ranges from 0 (no correlation, i.e., gametes within sub-populations are no more similar than gametes among sub-populations) to 1 (each sub-population is fixed with a different allele). F_{st} analysis has been used to identify regions of the genome associated with domestication and selective sweeps associated with breeding (Yan et al. 2017). It can also improve GS and genome-wide association studies (GWAS). For example, Chang et al. (2019) showed that prioritizing and weighting SNPs based on their F_{st} values can increase the accuracy of genomic predictions by more than 5%. Yan et al. (2017) in soybean found that combining GWAS and fixation index analysis helped identify QTLs for seed weight.

The goal of this study is to characterize the genome properties of two major oil palm breeding populations,

Deli and La Mé, focusing on key parameters for genomic predictions, namely LD, haplotype sharing, N_e , and F_{st} .

Materials and methods

Plant material and experimental design

The plant material used in this experiment consisted of individuals of the Deli and La Mé populations and their hybrid crosses. It comprised 943 genotyped individuals with 423 Deli, 140 La Mé, and 380 Deli × La Mé hybrid individuals. The Deli and La Mé populations used here were complex, involving several families with varying sizes and levels of relatedness. Thus, the Deli individuals belonged to 89 families of full-sibs with a mean size of 4.8 individuals (ranging from one to 60 individuals). The La Mé individuals belonged to 24 families of full-sibs with a mean size of 5.8 individuals (ranging from one to 31 individuals). Detailed pedigree information of these two populations is known over several generations (Cros et al. 2017). The Deli × La Mé hybrid individuals were obtained crossing 67 and 63 of these Deli and La Mé individuals, respectively, according to an incomplete factorial design. The hybrid individuals belonged to 101 crosses comprising on average 3.8 individuals (ranging from one to 10). For the construction of the genetic map, all the genotyped Deli, La Mé, and Deli × La Mé individuals were used, as well as the non-genotyped individuals comprised in their pedigree, for a total of 1788 individuals. For the other parts of the study, we only used the genotyped individuals of the Deli and La Mé breeding populations. The plant material was located partly in North Sumatra, on the SOCFINDO estate (Indonesia), and partly in Benin, on the INRAB research station of Pobè.

Genotypic data

Molecular data were obtained by genotyping by sequencing (GBS) (He et al. 2014). DNA extraction, GBS, and SNP calling were performed based on the procedure described in Cros et al. (2017). The sequence data were processed using Tassel GBS version 5.2.44 (Glaubitz et al. 2014). The reference genome of Singh et al. (2013) was used for alignment with Bowtie2 software (Langmead and Salzberg 2012). Biallelic SNPs were the only variants kept. SNP data points with depth below 10 were set to missing and only SNPs with less than 50% missing data in the two breeding populations were kept. SNPs with the sum of depth per datapoint above 550,000 and SNPs with 100% heterozygote genotypes were discarded. Individuals with more than 50% missing data were removed. Finally, we obtained 7324 SNP markers, common to both breeding populations. It included 5598 SNPs located on the anchored sequences of the genome (i.e.,

the 16 chromosomes of Singh et al. (2013) (Table 1). The average percentage of missing data per SNP was 11% in Deli and 13% in La Mé, with median values below 5% (Supplementary Figure 1).

Construction of the genetic map

The genetic map was made using LepMAP3 software version 0.4 (Rastas 2017). First, module ParentCall2 was used to call the parental oil palm genotypes, with parameters `removeNonInformative=1`, to remove the non-informative markers (monomorphic or homozygous in both parents), and `halfSibs=1`. Second, the Filtering2 module was used to remove SNPs segregating in a non-Mendelian fashion using `dataTolerance=0.001`. Third, the SeparateChromosomes2 module assigned markers into linkage groups (LGs) by computing all pairwise LOD scores between markers and joined markers with LOD scores higher than the user-given parameter `LodLimit`, which was set to eight. Fourth, the JoinSingles2All module assigned single markers to the existing LGs by computing LOD scores between every single marker and markers from the existing LGs, using `lodLimit=4` and `iterate=1`. Finally, OrderMarkers2 ordered the markers within each LG by maximizing the likelihood of the data given the order and using the Kosambi mapping function for the conversion of recombination frequencies into map distances (centiMorgan, cM) (Rastas 2017). To join the maps of both male and female parents, the `sexAveraged` argument was set to 1. The individuals that were associated with outlier values

in terms of the number of crossing-overs were identified in preliminary analysis and removed before the map construction. The markers which created large gaps at the top or bottom part of the LGs were discarded according to software developer recommendations. This resulted in a genetic map where the 16 largest LGs had largely higher numbers of SNPs than the remaining LGs, which were discarded to keep a genetic map with the number of LGs corresponding to the number of chromosomes of oil palm.

Comparison of genetic and physical maps

The genetic map and physical map, showing the positions on the reference genome of Singh et al. (2013), were visualized using the R package LinkageMapView (Ouellette et al. 2018). We used MareyMap (Siberchicot et al. 2017) to plot the genetic position of the molecular markers against their physical position.

Within population linkage disequilibrium and persistence between populations

Analyses of linkage disequilibrium (LD) were performed in each breeding population using the PLINK software (Purcell et al. 2007). It computed pairwise estimates of LD by the classical measure of the squared correlation of allele frequencies at diallelic loci (r^2) and r . Before the computation of the r^2 , the missing data points in the Deli and La Mé individuals were imputed using Beagle5.2 (Browning et al.

Table 1 Summary of the physical map (SNPs located on the assembled part of the genome)

Chromosome name	Number of markers	Length (bp)	Average distance of markers (bp)	Maximum distance of markers (bp)	Minimum distance of markers (bp)
EG51_1	271	65,071,148	2,409,88.92	4,189,850	1
EG51_2	311	63,345,076	202,765.00	5,820,505	1
EG51_3	318	58,158,439	182,741.28	2,511,138	1
EG51_4	257	42,716,717	163,396.59	7,354,398	1
EG51_5	222	55,995,026	250,540.60	4,155,834	1
EG51_6	154	43,622,229	282,049.04	5,714,930	1
EG51_7	219	51,181,318	232,528.43	3,220,718	1
EG51_8	162	31,376,194	194,283.67	1,759,800	1
EG51_9	79	21,017,043	269,303.63	5,020,104	1
EG51_10	154	39,935,972	260,564.44	2,279,224	1
EG51_11	133	28,384,088	198,092.74	2,810,164	1
EG51_12	132	30,035,350	192,305.95	4,702,868	1
EG51_13	90	37,835,912	418,385.07	4,660,806	1
EG51_14	122	23,067,684	187,621.79	2,011,138	1
EG51_15	66	25,884,061	393,063.51	3,063,683	1
EG51_16	92	23,929,541	237,246.78	1,759,080	1
Sum	2782	641,555,798			
Mean	173.87	40,097,237.38			

2018), independently for each breeding population. For the SNPs located on the assembled parts of the genome, the r^2 values between pairs of SNPs were plotted against physical distances (Mbp). For the SNPs located on the genetic map, the r^2 values were plotted against genetic distances (cM). The LD decay was plotted up to a 0.8-Mbp distance for physical positions and 3 cM for genetic positions. The relation between the r^2 values and distances was modeled by fitting local polynomials with the functions “locpoly” and “dpill” of the R package KernSmooth 2.23 (Wand 1995), as done for example in Yamamoto et al. (2016).

The persistence of LD between populations (De Roos et al. 2009) was measured by the correlation of the r measure of LD between populations given by PLINK (r_{LD}). r_{LD} was computed between the two populations on the SNPs comprised in windows defined along with the genetic and

physical maps, over a distance up to 90 cM and 50 Mbp, respectively. r_{LD} values can vary from -1 to 1 , with a value close to 1 indicating a similar LD pattern in the two populations for the SNPs located in the genomic window considered.

Haplotype sharing

This analysis was done with the SNP data phased with Beagle5.2 (Browning et al. 2018). Sliding windows were defined along the chromosomes and linkage groups, with an overlap of 50%. Fifteen window sizes were used for physical distances, from 10 Mbp to 100 bp, and seven window sizes were used for genetic distances, from 10 to 0.01 cM. The window sizes were considered by decreasing order and, for each window of a given window size, the list

Fig. 1 Distribution of minor allele frequency (MAF) in Deli and La Mé oil palm breeding populations

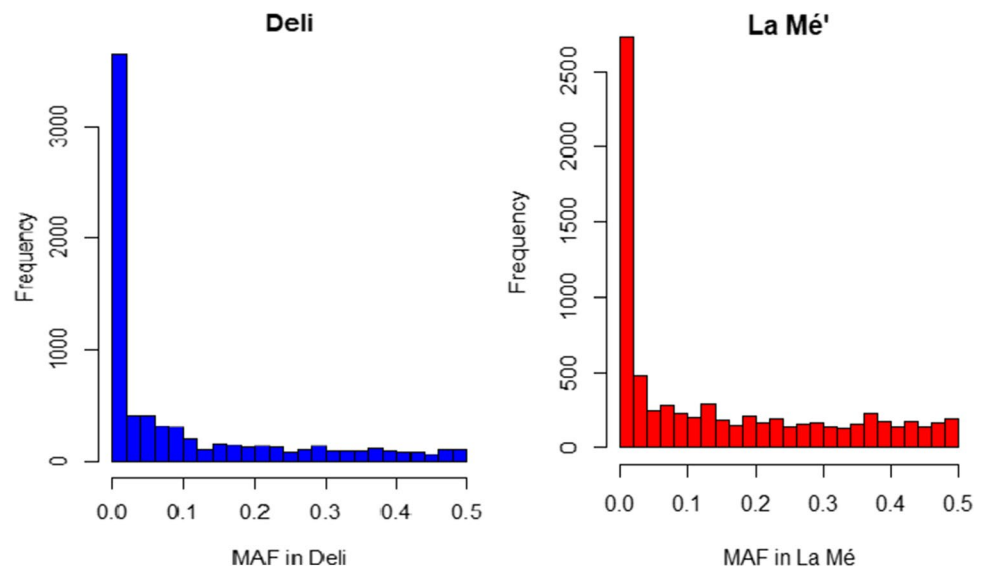
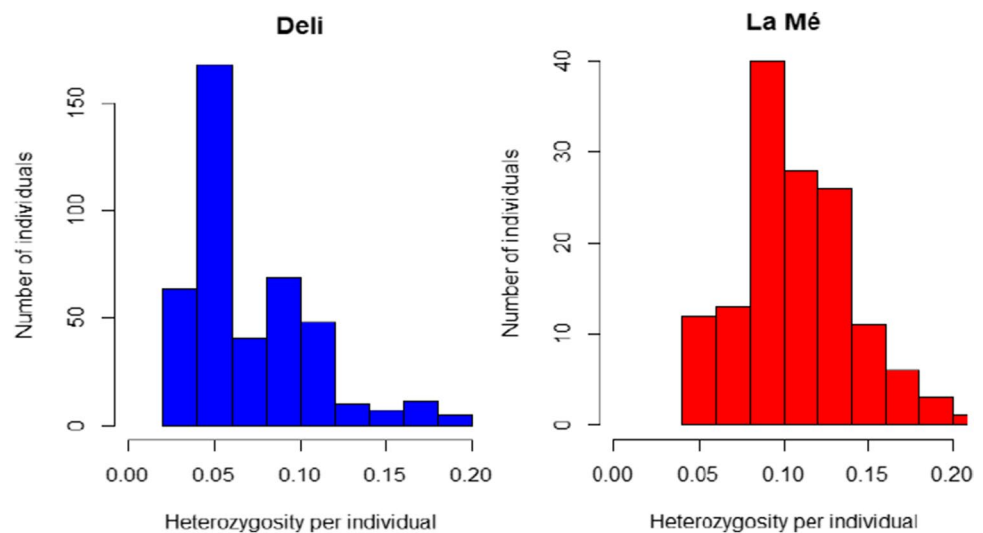


Fig. 2 Distribution of the percentage of heterozygosity per individual for Deli and La Mé oil palm breeding populations



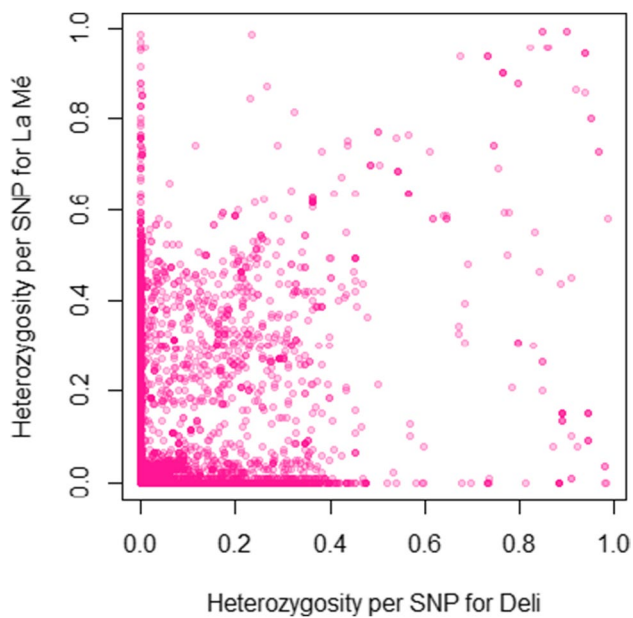


Fig. 3 Correlation of heterozygosity per SNPs among Deli and La Mé oil palm breeding populations. Each dot represents an SNP. Color intensity indicates density of overlapping dots

of haplotypes existing in each population was made after discarding the haplotypes with an actual length shorter than the next window size. In the end, to avoid redundancy that could result from the overlap between windows, only a single copy of the duplicated haplotypes (i.e., haplotypes identical in sequence and starting at the same position) was kept. Finally, the length of the haplotypes, the percentage of haplotypes common to the two populations, and, for the common haplotypes, their frequency in each population were computed. This analysis was done using a custom R script.

Effective size

The effective size was estimated with the LD method of Waples and Do (2008) implemented in the NeEstimator 2.1 software (Do et al. 2014). The computation was made separately in each population using the SNPs located on the genetic map and selecting the “random mating” option of the software. The confidence interval of N_e values was obtained by the Jackknife method on samples.

Fixation index

Pairwise F_{st} between Deli and La Mé was estimated according to Wright (1931), using the 7324 or 5598 SNPs available with $MAF > 1\%$ and subsets of 100 random individuals

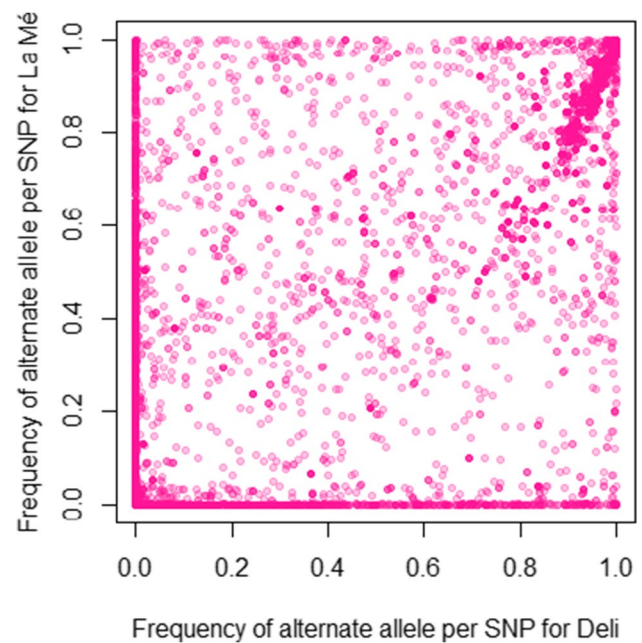


Fig. 4 Correlation of frequency of alternate allele per SNP among Deli and La Mé oil palm breeding populations. Each dot represents an SNP. Color intensity indicates density of overlapping dots

per population, to avoid a bias in computing the F_{st} values between an unequal number of genotyped individuals per population (Gondro et al. 2013). The F_{st} was obtained using the SNPRelate R package (Zheng et al. 2012).

Results

Allele and genotype frequencies

The distribution of minor allele frequency (MAF) in Deli and La Mé oil palm populations showed a reduction in the number of SNPs with the increase of MAF (Fig. 1). The average MAF was 0.09 for Deli and 0.14 for La Mé. In both populations, most SNPs had low MAF values. Thus, the percentage of SNPs with $MAF < 0.05$ was 60.5% in Deli and 49.7% in La Mé.

The percentage of heterozygosity per individual ranged from 1.9% (Deli) to 20.9% (La Mé) (Fig. 2). Deli was the population with the lowest percentage of heterozygosity (mean 7%, versus 10% for La Mé).

The correlation of heterozygosity per SNPs between the two populations (Fig. 3) showed that the majority of SNPs were, in one population, fixed or almost fixed (i.e., concentrated alongside the x and y axes) while, in the other population, they had a much larger level of heterozygosity.

Similarly, the correlation in the frequency of alternate alleles per SNP between populations demonstrated that most SNPs have distinct segregation patterns among populations,

with SNPs largely concentrated alongside the x and y axes (Fig. 4). A large proportion of SNPs thus appeared fixed or almost fixed with the reference allele in one population (i.e., frequency of alternate allele equal or close to 0), while having a significant proportion of the alternate allele in the other population.

High-density genetic map

The process of construction of the genetic map yielded a set of 16 linkage groups (LGs). The genetic map comprised 4252 SNPs, spread over 2782 unique positions (Table 2 and Fig. 5, Supplementary Figure 2), and spanned 1778.52 cM. Even coverage of the genome was achieved, with an average mapping interval between adjacent positions of 0.67 cM and the largest gaps between adjacent positions ranging from 3.31 cM (LG11) to 6.66 cM (LG14). The size of the LGs ranged from 215.72 cM (LG1) to 64.75 cM (LG16) (Table 2). The number of unique SNP positions mapped to each linkage group ranged from 87 (LG 14) to 358 (LG1), with a mean of 174.93 per linkage group. The biggest gap size between SNPs ranged from 3.31 cM (LG11) to 6.66 cM (LG14).

Comparison of genetic and physical maps

The physical and genetic orders were in general agreement, with a Spearman rank correlation above 0.7 for 15 LGs out

of 16 (Table 2 and Fig. 6). However, upturns of large chromosome segments between the genetic map and the reference genome existed in a few cases, for example in chromosome EG51_16 (Fig. 6). Also, punctual disagreements between physical and genetic distances concerning a few SNPs appearing as outliers, i.e., far apart from the regression line, were observed in most chromosomes (Fig. 7).

The recombination rate was 2.85 cM/Mbp on average, ranging from 1.78 cM/Mbp (LG15) to 3.87 cM/Mbp (LG13) (Table 2).

Within-population linkage disequilibrium and persistence between populations

The decay of LD between pairs of SNPs according to the genetic distances is shown in Fig. 8. The LD reached high values (>0.6) for short distances between SNPs. It was higher in Deli than that in La Mé for all distances. For example, considering the r^2 value of 0.3, the corresponding distance between SNPs was 1.05 cM in Deli and 0.9 cM in La Mé (Fig. 8). The difference between the two populations was small for short distances and increased with the distance between markers. Similar trends were observed when plotting LD against physical distances (Fig. 9), although the r^2 values reached higher levels (i.e., around 0.80), as a consequence of the higher number of markers on the physical map than on the genetic map. The distance corresponding to $r^2=0.3$ was 0.22 Mbp in Deli and 0.21 Mbp in La Mé.

Table 2 Summary of the genetic map and comparison with the physical map

Linkage group	Number of markers	Length in cM	Average gap size (cM)	Biggest gap size (cM)	Number of unique positions	Corresponding chromosome (Singh et al. 2013)	Number of common markers	Spearman correlation (absolute value)	Recombination rate (cM/Mb)
LG1	554	215.72	0.60	5.20	358	EG51_2	271	0.86	2.19
LG2	436	142.59	0.51	6.42	279	EG51_1	311	0.83	3.41
LG3	432	155.39	0.50	4.75	309	EG51_3	318	0.80	2.67
LG4	326	129.51	0.60	4.91	218	EG51_7	257	0.95	2.53
LG5	312	142.82	0.64	5.09	223	EG51_4	222	0.72	3.34
LG6	278	111.51	0.68	4.35	164	EG51_6	154	0.94	2.56
LG7	277	142.75	0.69	5.04	207	EG51_5	219	0.94	2.55
LG8	220	94.21	0.66	5.70	144	EG51_10	162	0.79	2.36
LG9	225	88.64	0.88	6.04	102	EG51_16	79	0.54	3.70
LG10	216	113.85	0.76	4.92	150	EG51_8	154	0.91	3.63
LG11	204	90.64	0.63	3.31	144	EG51_12	133	0.71	3.02
LG12	185	65.27	0.54	3.80	123	EG51_11	132	0.97	2.30
LG13	163	81.31	0.84	4.95	98	EG51_9	90	0.86	3.87
LG14	158	72.31	0.84	6.66	87	EG51_14	122	0.90	3.13
LG15	136	67.25	0.68	4.40	100	EG51_13	66	0.93	1.78
LG16	130	64.75	0.70	4.54	93	EG51_15	92	0.96	2.50
Sum	4252	1778.52			2799		2782		
Mean	265.75	111.15	0.67	5.00	174.93		173.875	0.85	2.85

Genetic map of Deil and LaMè oil palm populations

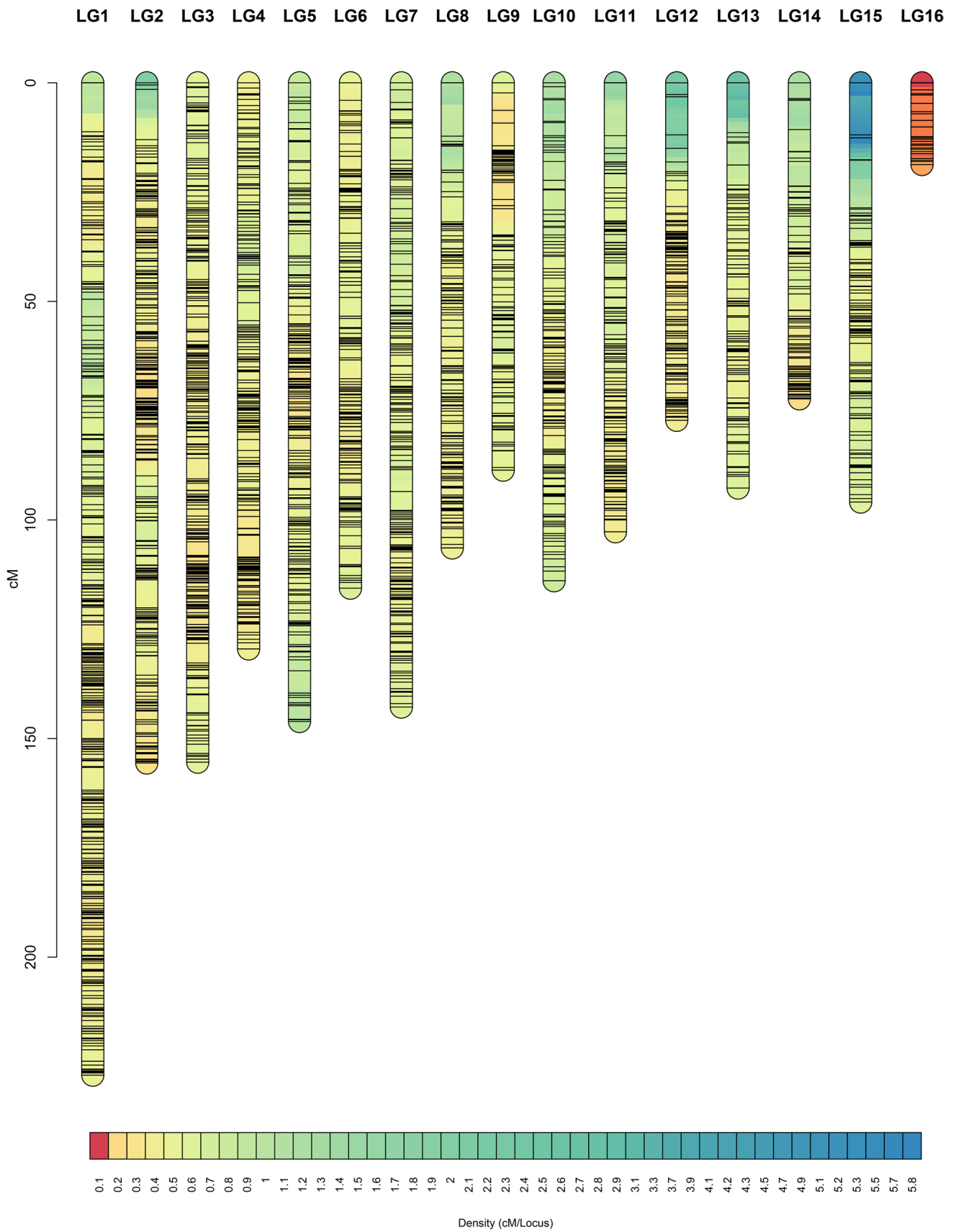


Fig. 5 Genetic map of Deli and La Mé oil palm populations with 4252 SNP markers. The colors indicate the density of markers according to the bottom scale (cM/locus).

A high correlation of r values between populations (r_{LD}) was observed for close markers, i.e., r_{LD} above 0.6 for SNPs separated by a distance <0.5 cM on the genetic map or <1 kbp on the physical map (Fig. 10). The r_{LD} value decreased sharply with the distance between SNPs, and was thus divided by two before 2 cM and 5 Mbp, and became negligible at distances above 50 cM or 50 Mbp.

Haplotype sharing

The percentage of shared haplotypes between Deli and La Mé populations according to the length of the genomic window is represented in Figs. 11 and 12. A large proportion of haplotypes were common between pairs of populations when considering short distances. Thus, 50% of the haplotypes with length around 30 bp (Fig. 11) and 40% of the haplotypes with length around 3600 bp were common to the two populations, and 40% of the haplotypes with length around 0.20 cM were common to the two populations (Fig. 12). As expected, when the length of the haplotypes increased, the percentage of shared haplotypes between populations decreased. The decrease was fast, with the percentage of common haplotypes falling below 20% for haplotypes longer than 300 kbp and 2.5 cM.

The frequency of the common haplotypes coincided to some extent for short haplotypes, while the differences increased for longer haplotypes. Thus, among the common haplotypes identified with a window size of 100 bp, more than one-half (51.6%) of the ones with a frequency $>90\%$ in Deli also had a frequency $>90\%$ in La Mé. This value fell to 25% for haplotypes identified with a window size of 50 kbp and to 14% for a window size of 500 kbp.

Effective size

The two populations had small N_e values, i.e., 3 for Deli (95% confidence 2.7-3.3) and 3.6 for La Mé (3.0-5.2).

Fixation index

The F_{st} between Deli and La Mé was 0.53. Supplementary Figure 3 showed the F_{st} between the two populations at the chromosome level. Depending on the region of the genome considered, there were large variations in the F_{st} among the two pairs of populations. Thus, several regions of the genome had high F_{st} values (>0.6), in particular on chromosomes EG51_2, EG51_8, and EG51_13.

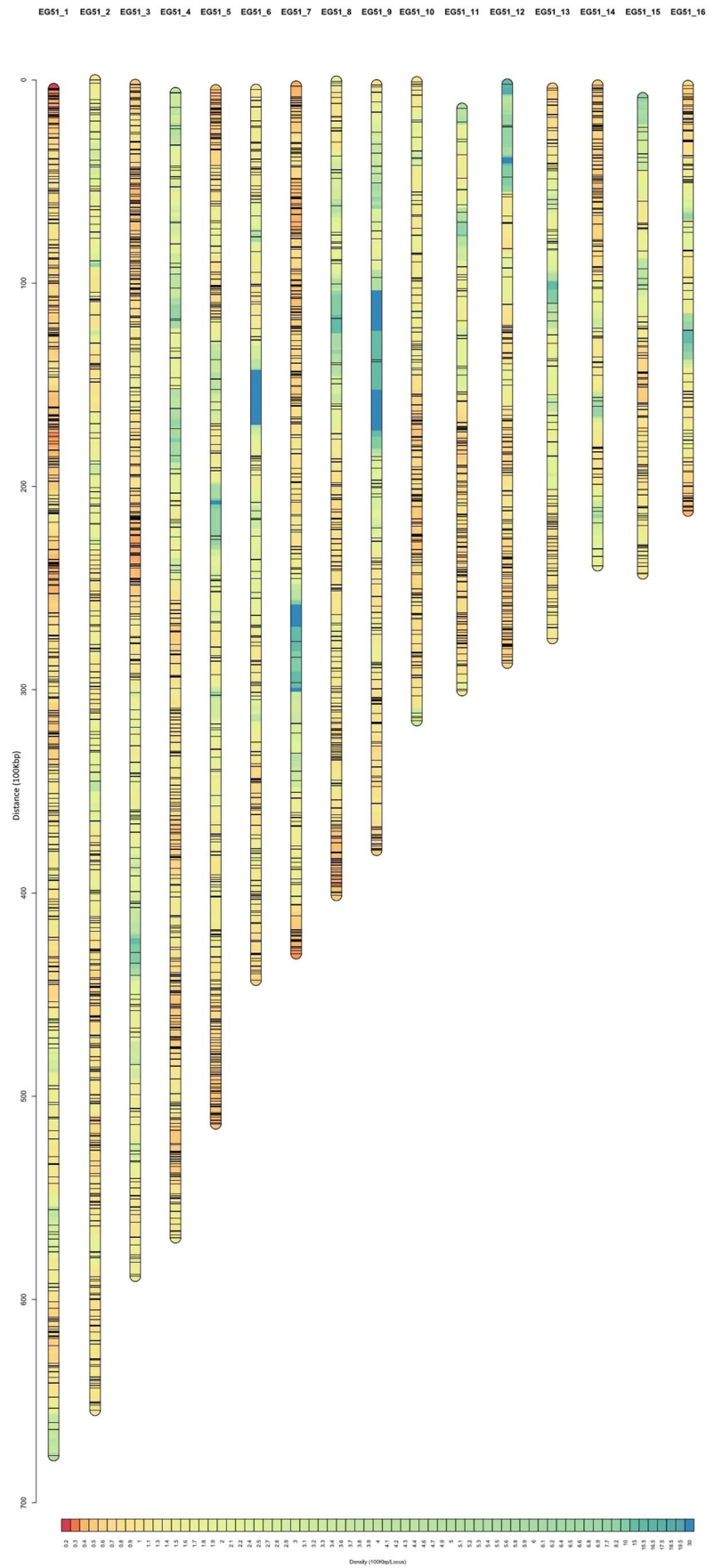
Discussion

In this paper, we characterized the genome properties of two key oil palm breeding populations, Deli and La Mé, using SNP data obtained by GBS. This genotyping approach, despite the higher rate of missing data and errors than SNP array and SSR, appears relevant for the present study, as indicated by previous articles on other species where genetic maps and LD profiles were obtained from GBS, for instance in coconut (Rajesh et al. 2021), hevea (de Souza et al. 2018), and tea (Niu et al. 2019). Also, the Lep-MAP3 software used here to generate the genetic map is particularly robust against missing data, as it does not rely on two-point analysis, and against genotypes of lower reliability due to low coverage sequencing (Rastas 2017).

Within-population linkage disequilibrium and persistence between populations

The pattern of LD is one of the utmost factors affecting both GWAS and GS since both methods rely on LD between markers and causal polymorphisms (Sorkheh et al. 2008; Hayes et al. 2009; Yadav et al. 2021). LD is thus one of the major factors that determine the number of markers required (Heffner et al. 2009; Lebedev et al. 2020). r^2 values of 0.3 are considered a minimum to get reliable results in GS studies and GWAS (Bejarano et al. 2018). Here, when considering the genetic distances, the r^2 value reached 0.3 with SNPs separated by around 1.05 cM in Deli and 0.9 cM in La Mé (Fig. 8). As our genetic map spanned 1778.52 cM, achieving this distance between adjacent SNPs requires around 1700 SNPs for Deli and 2000 SNPs for La Mé. When considering the physical distances, the r^2 value of 0.3 was achieved with SNPs separated by around 220 kbp in Deli and 210 kbp in La Mé (Fig. 9). As here the genome length covered by SNPs spanned 643 Mbp, achieving these distances between adjacent SNPs would take around 2900 SNPs in Deli and 3100 SNPs in La Mé, which can be considered close to the value obtained from the LD decay along with the genetic map. Considering that the goal should be to cover the whole genome and that the oil palm genome spans 1.8 gigabases (Singh et al. 2013), 10,000 SNPs would be enough to reach the r^2 value of 0.3 in the two populations studied here (as this corresponds to around 8200 SNPs in Deli and 8600 La Mé). The effect of marker density on the GS accuracy has already been evaluated on oil palm datasets comprising the populations considered here. It showed that, depending on the study and trait, the number of SNPs required to achieve maximum GS accuracy was found to range from 500 to 7000 (Cros et al. 2017; Nyouma et al. 2020). This is in agreement with the results obtained from the LD analysis.

Fig. 6 Physical map of Deli and La Mé oil palm populations with 5598 SNP markers. The colors indicate the density of markers according to the bottom scale (100 kbp/locus)



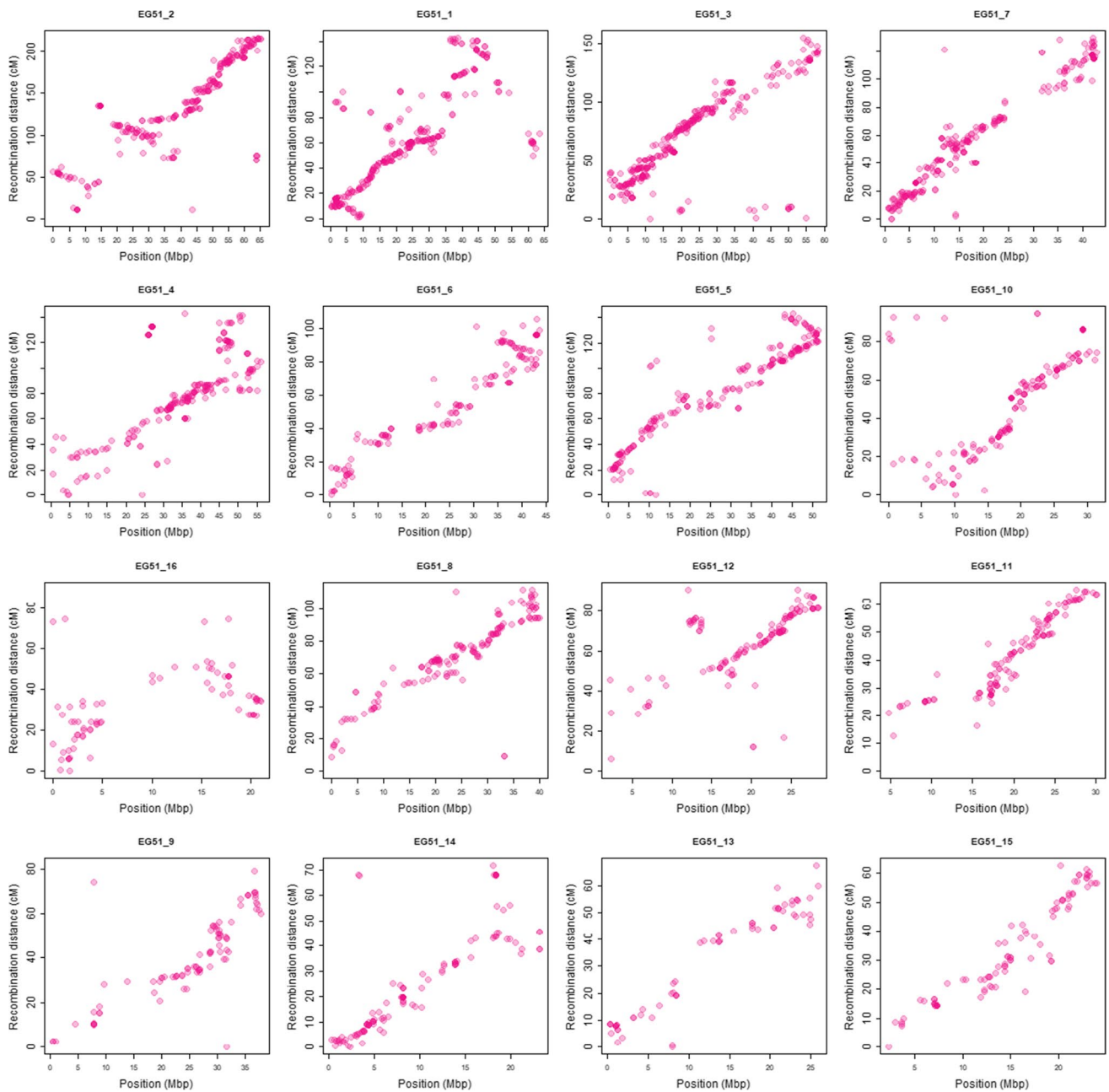


Fig. 7 Visualization of marker genetic positions (cM) versus physical positions (Mbp) for each chromosome. The plots are ordered according to linkage groups number. Each dot represents an SNP. Color intensity indicates density of overlapping dots

Our results also revealed that the speed and the magnitude of LD decay varied between the breeding populations. The two populations were submitted to a **founding bottleneck of similar magnitude**. A bottleneck increases LD and slows down the LD decline (Tenaillon et al. 2008). We can assume the higher value of LD in the Deli population in all genomic distances resulted from the fact that its history was marked by a larger number of **generations of selection and inbreeding** than in La Mé, with the bottleneck event in the Deli history dating back to 1848 against the 1920s in La Mé.

High correlations of r values between populations ($r_{LD} > 0.6$, corresponding to $r_{LD}^2 > 0.25$) were obtained considering the markers that were the closest from each other, i.e., with distances < 0.5 cM on the genetic map or < 1 kbp on the physical map. Similarly, a large proportion of haplotypes was common between Deli and La Mé when considering windows of reduced size, with $> 40\%$ of haplotypes with lengths below around 3600 bp or 0.20 cM being common in the two populations. This explains the results of Nyouma et al. (2020, 2022), who found, using the same breeding

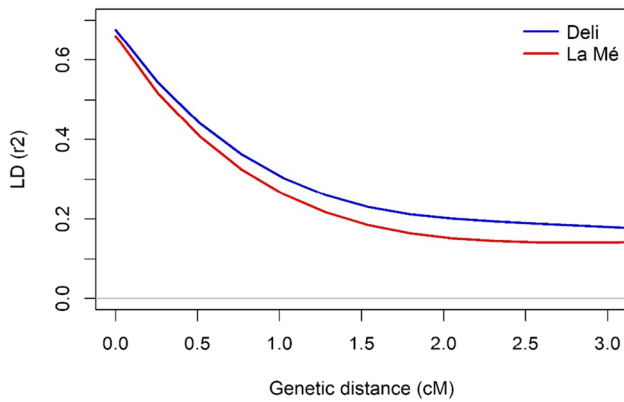


Fig. 8 Average genome-wide pattern of linkage disequilibrium (LD) decay between pairs of SNPs (r^2) according to the genetic distance (cM) between SNPs, for Deli and La Mé oil palm breeding populations

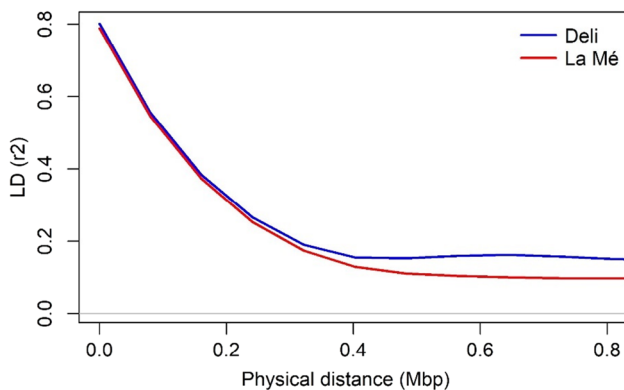
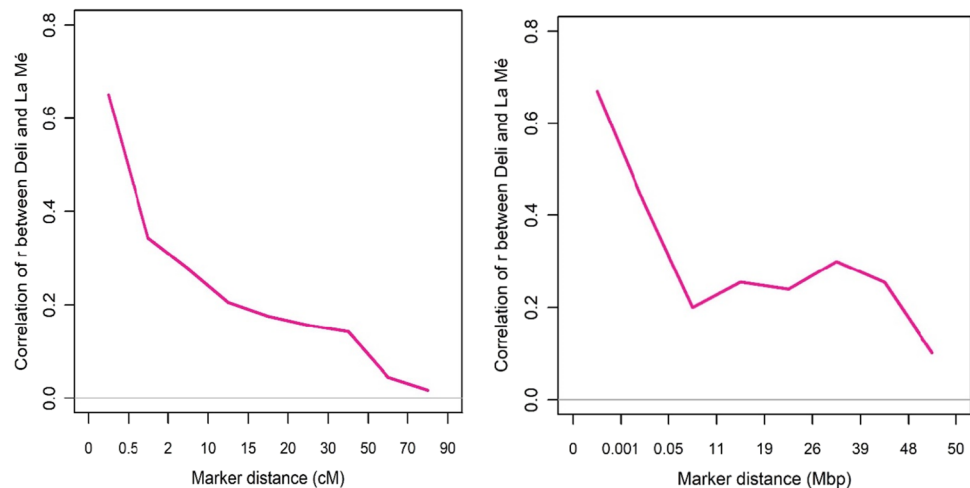


Fig. 9 Average genome-wide pattern of linkage disequilibrium (LD) decay between pairs of SNPs (r^2) according to the physical distance (Mbp) between SNPs, for Deli and La Mé oil palm breeding populations

Fig. 10 Correlation of the r measure of LD between populations as a function of genomic distance in cM (right) and Mbp (left)



populations and the same genotyping approach (GBS), that for GS predictions in oil palm it was better not to model the parental origin of marker alleles. The superiority of GS models ignoring the parental origin of marker alleles over models considering it does not imply a complete persistence of phases between markers and QTLs among populations. Indeed, models that consider the parental origin of marker alleles are more complex and require the estimation of more parameters, possibly reducing their predictive ability, despite their ability to better depict the genetic differences between the population. The current study and the previous results of Nyouma et al. (2020, 2022) indicate that the level of conservation of phases among the Deli and La Mé populations captured with the present marker density is high enough to favor models ignoring the parental origin of marker alleles. A similar conclusion was reached by Technow et al. (2012) in maize, to explain the cases where this type of model outperformed the population-specific allele models. To further investigate this aspect, it would be interesting to study the correlation of marker effects obtained by GS models between Deli and La Mé populations, as done for maize in Technow et al. (2014). To our knowledge, this is the first study investigating the persistence of LD and phases between oil palm populations.

Other studies investigated the pattern of LD in oil palm, in particular Teh et al. (2016) and Kwong et al. (2016), using high-density SNP arrays. However, the results are difficult to compare, as the studies involved different populations, in particular inter-group hybrids, against parental populations in our study. However, Kwong et al. (2016) included in their work two breeding populations, JL×DA and GM×DA, that were mostly of Deli origin. Their LD value decreased by 50% from around 25 to 200 kbp, i.e., in the same range as the value found in our study (around 175 kbp). A previous study considered the same breeding populations as in the present study but used SSR markers (Cochard 2008). The results were however in agreement, with Deli having the

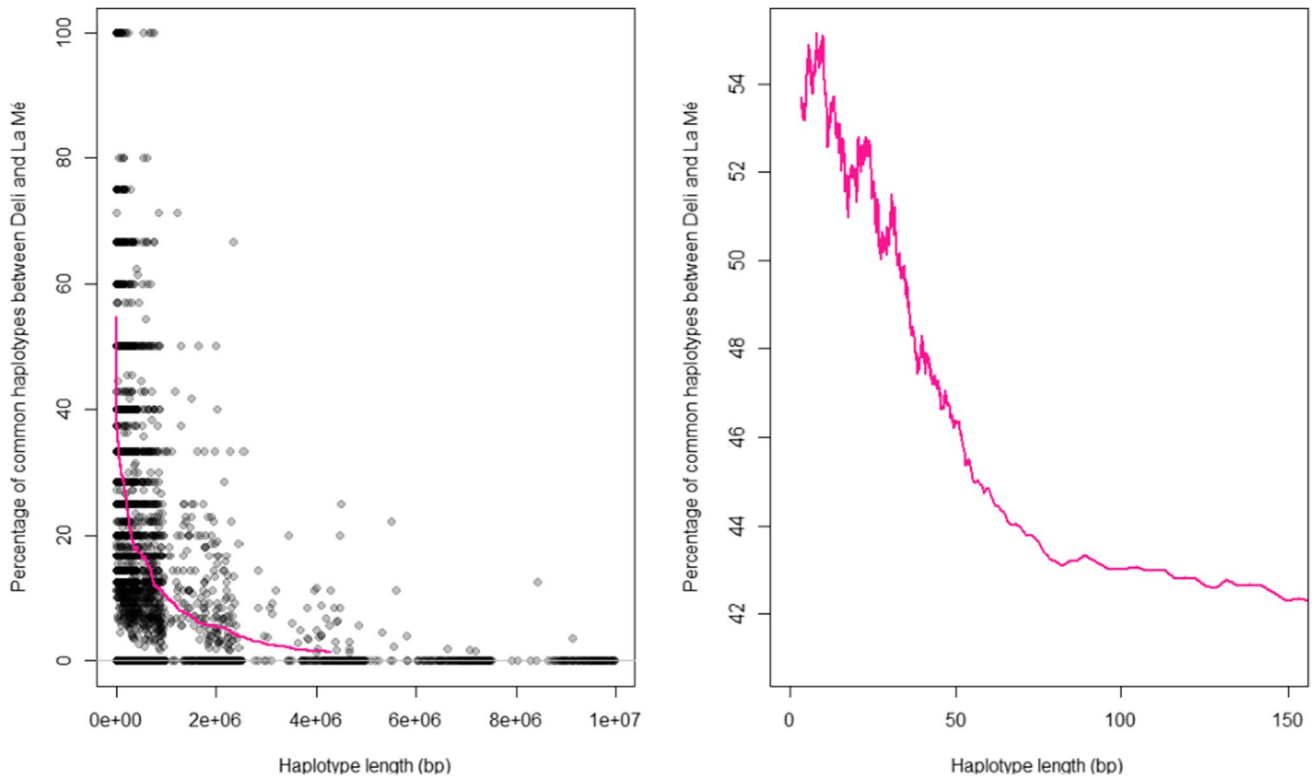


Fig. 11 Percentage of unique haplotypes shared between Deli and La Mé oil palm breeding populations according to the haplotype length in bp. Each dot represents a haplotype. Color intensity indicates density of overlapping dots. The smoothing curve in pink is the rolling average

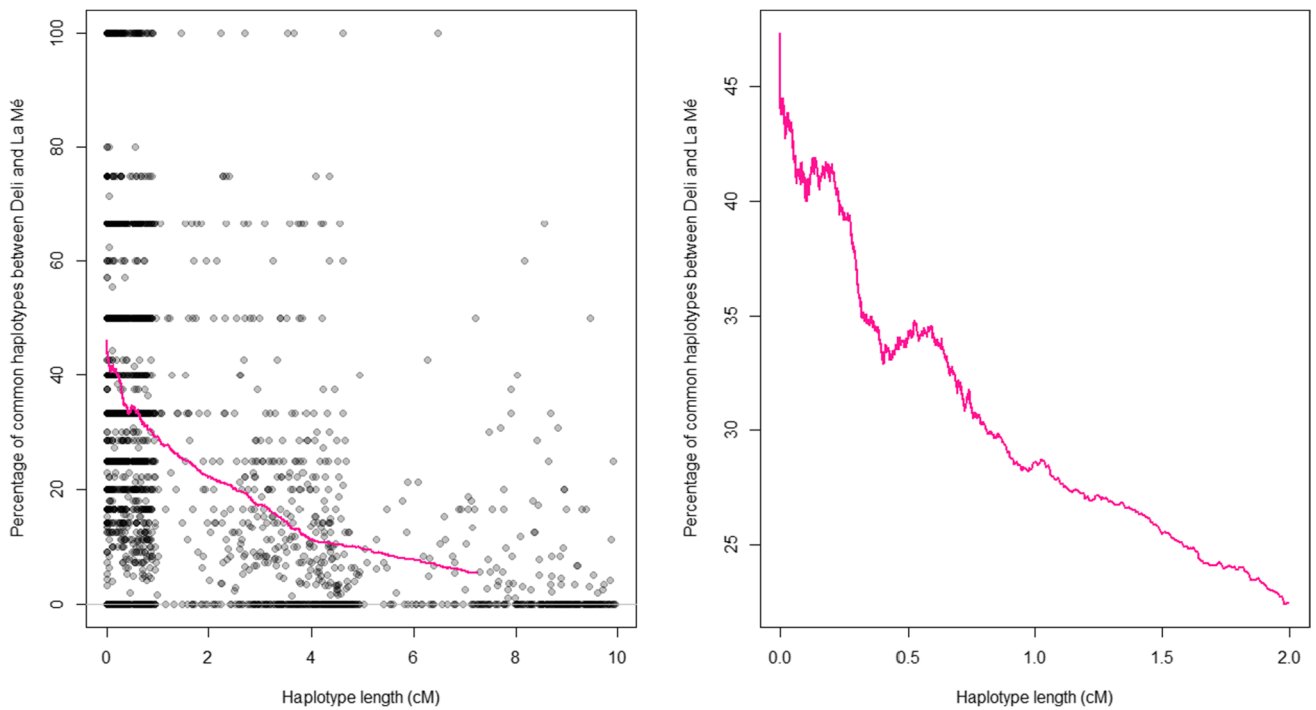


Fig. 12 Percentage of unique haplotypes shared between Deli and La Mé oil palm breeding populations according to the haplotype length in cM. Each dot represents a haplotype. Color intensity indicates density of overlapping dots. The smoothing curve in pink is the rolling average

highest LD values. The consistency of these results shows that GBS is a suitable approach for LD studies, despite a higher rate of missing values (Supplementary Figure 1) and genotyping errors compared to SNP arrays and SSR, while providing much higher marker density than SSRs.

Genetic differentiation between Deli and La Mé

The F_{st} study, the correlation of heterozygosity per SNP, the correlation of frequency of alternate allele per SNP, and the decrease of persistence of LD and of haplotype sharing with increasing SNP distance showed a significant degree of differentiation among the two oil palm breeding populations. Reciprocal recurrent selection (RRS) certainly contributed to this differentiation, as suggested by results obtained in the heterotic groups used in maize RRS. Thus, the Iowa stiff stalk and corn borer synthetic populations diverged along with RRS cycles, with the fixation of different alleles and a steady increase in F_{st} (Labate et al. 1999; Gerke et al. 2015), and large differences in allele frequencies were also found between Dent and Flint populations (Technow et al. 2014). Our results are also in agreement with the oil palm study of Cochard et al. (2009), who concluded that the Deli population derived from a group comprising Benin, Nigeria, Cameroon, Congo, and Angola populations, while the populations west of Benin were genetically more different from those of Deli. This supports the idea that the four founders of the Deli population were collected in Central Africa rather than in West Africa (Cochard et al. 2009). The variation found in the F_{st} profile, which reached high values (>0.6) in some genomic regions, suggests that F_{st} is likely to be of interest for studying signatures of selection. This could help identify candidate genes, especially for traits with contrasting phenotypic values between breeding populations, such as bunch number and bunch weight between A and B groups. However, a higher SNP density seems necessary to obtain clearer profiles with more pronounced peaks (Porto-Neto et al. 2013) that could be linked to genes of interest based on the available information on oil palm annotation.

Effective size

To our knowledge, there was so far no estimate available of N_e for the La Mé breeding population. The small values obtained here for the Deli and La Mé populations are not surprising given their history, with a small number of founders and under the effect of inbreeding. In Cros et al. (2014), N_e was estimated for a subset of 104 Deli individuals from the population used here, with 16 SSR markers chosen on different linkage groups and the LD method of Waples and Do (2008). This gave a N_e of 5 ± 1.1 (SD), i.e., similar to the result we obtained here. This indicates the robustness of the method against marker type and density.

The small N_e values obtained here also explain the fact that GS can be implemented with small training populations and low marker density. Thus, in previous studies, GS models trained with data from only 108 Deli and 102 La Mé individuals were efficient enough to replace phenotypic selection before clonal trials (Nyouma et al. 2020) while GS accuracy plateaued with only 500 to 2000 SNPs, depending on the trait (Cros et al. 2017).

Comparison of genetic and physical maps

The construction of genetic linkage maps using SNP markers is common in oil palm (see, for instance, Jeennor and Volkaert 2014; Ting et al. 2014; Lee et al. 2015; Bai et al. 2018a, b; Gan et al. 2018; Ong et al. 2019; Herrero et al. 2020). The genetic linkage maps helped identify genomic regions having major genes and QTLs that control oil yield (Montoya et al. 2013; Jeennor and Volkaert 2014; Tisné et al. 2015), palm oil fatty acid composition (Singh et al. 2009; Montoya et al. 2013), vegetative growth (Ukoskit et al. 2014; Lee et al. 2015; Bai et al. 2018b; Teh et al. 2020), and resistance to diseases (Tisné et al. 2017; Daval et al. 2021). High-density maps were also used to improve the assembly of previously published genome sequences by assigning scaffolds originally unplaced (Ong et al. 2019, 2020). To our knowledge, the present study involved the largest number of individuals genotyped for the construction of a genetic map in oil palm. Another original aspect of our genetic map is the use of complex plant material including several families with varying degrees of relatedness, several generations, and different populations. In contrast, the previously published oil palm genetic maps were usually constructed from full-sib families (e.g., Watson et al. 2001; Cochard et al. 2009; Ting et al. 2013; Ukoskit et al. 2014), although Billotte et al. (2010) used a factorial design. To our knowledge, only Cochard et al. (2015) and Daval et al. (2021) constructed genetic maps from populations with similar levels of complexity. However, they used SSR markers and the CRI-MAP software (Green et al. 1990), which seems less efficient than LepMAP3, as it has problems handling large pedigrees with large numbers of bi-allelic markers particularly when there are lots of missing parental and grandparental genotypes. The map of our study is shorter than the map of Cochard et al. (2015), which reached 1935 cM and was obtained using a similar oil palm population. This might be a consequence of the marker type, as it was shown that SSRs led to inflated maps compared to SNPs (Ball et al. 2010).

The linkage map presented here, with an average marker density of one SNP in every 0.67 cM when considering unique positions, had a denser genome coverage compared to most previously published SNP oil palm

genetic linkage maps, like Ting et al. (2014), with one marker in every 1.40 cM and Pootakham et al. (2015) with one marker in every 1.26 cM. However, our map is less dense than the genetic linkage maps constructed by Ong et al. (2019, 2020), with one marker in every 0.04 cM, 0.05 cM, and 0.18 cM, depending on the map, and Bai et al. (2018a), with one marker every 0.29 cM, and Herero et al. (2020), with one marker in every 0.57 cM. Most of these variations in terms of the marker density of the genetic maps can be explained by differences in genotyping approaches and the size of the populations (Ferreira et al. 2006; Semagn et al. 2006; Seyum et al. 2021). Combining high-throughput genotyping and populations with at least 150 individuals appears as an efficient strategy to maximize marker density, as in Ong et al. (2019, 2020), Bai et al. (2018a), and the present study.

There were several upturns between the genetic and physical maps (Fig. 7). For example, LG 1, 2, 5, and 16 had large upturns for regions of the genome of more than 10 Mbp. Aside from potential genome assembly artifacts, this can be the consequence of genomic rearrangements between populations, as the reference genome (Eg5.1) was obtained on an individual of the AVROS oil palm population (Singh et al. 2013), which thus differed from the populations used for the genetic mapping (Deli and La Mé). To further investigate this aspect, we compared the position of our SNPs on Eg5.1 with their position on EgPMv6, a new version of Eg5.1 improved through the use of a high-density linkage map (Ong et al. 2020), but that was made available after the beginning of the present study (Supplementary Figure 4). We found that, although some upturns existed (in particular for the smallest chromosomes), the positions on the two genomes are in general agreement. Consequently, there are still disagreements between the genetic positions obtained here and the physical positions, even considering the improved assembly. For example, LG 2 had 100% of its SNP located on the same chromosome according to Eg5.1 and EgPMv6 (Eg5.1_1 and GK000077.1, respectively), and almost identical SNP order between the two assemblies, while large upturns existed between the genetic and physical positions (Fig. 7). This aspect deserves further study, which could be done using population-specific genetic maps and reference genomes. This requires new data, with more genotyped individuals per population and new reference genomes.

Conclusion

The present study focused on two key populations used for hybrid breeding in oil palm, Deli and La Mé, and estimated genetic parameters affecting GS accuracy. A high-density genetic map was constructed from a complex population including several families with varying sizes

and levels of relatedness and with different genetic backgrounds. It included 4252 SNPs from GBS and spanned 1778.52 cM, with an average recombination rate of 2.85 cM/Mbp. The LD at $r^2=0.3$, considered as the minimum to get reliable results for genomic predictions, spanned over 1.05 cM/0.22 Mbp in Deli and 0.9 cM/0.21 Mbp in La Mé. In the two populations, 10,000 SNPs would be enough to reach this level of LD. A high correlation of r values of LD between populations ($r_{LD}>0.6$) was obtained considering the markers separated by short distances, i.e., <0.5 cM on the genetic map or <1 kbp on the physical map. The percentage of common haplotypes was above 40% for short haplotypes (3600 bp or 0.20 cM). This resemblance decreased with the distance between SNPs, with for example the percentage of common haplotypes falling below 20% for haplotypes longer than 300 kbp. The F_{st} was high (0.53). Overall, the results showed strong genetic differentiation between Deli and La Mé, but the level of resemblance between them over short genomic distances likely explains the superiority of GS models ignoring the parental origin of marker alleles over models taking this information into account. The N_e values of the two populations were small (<5). Population-specific genetic maps and reference genomes would be of interest for future studies.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s13353-022-00708-w>.

Acknowledgements The authors acknowledge the GENES program of the Intra-Africa Academic Mobility Scheme of the European Union for financial support (EU-GENES:2017-2552/001-001). The authors acknowledge SOCFINDO (Indonesia), CRAPP (Benin), and PalmElit (France) for authorizing the use of the data for this study. We thank the UMR AGAP genotyping technology platform (CIRAD, Montpellier) and the CIRAD-UMR AGAP HPC data center of the South Green Bioinformatics platform (<http://www.southgreen.fr/>) for their help.

Author contribution EGS, DC, NHB, and JMB participated in the design of the study. EGS and DC performed the statistical analysis and wrote the manuscript. WGA, NHB, BC, FJ, and JMB contributed to the manuscript. PR and DL participated in the construction of the genetic linkage map. DA, HD, and BC participated in designing field experiments, producing the plant material, and managing field trials. The molecular data were generated by VR and VP. All the authors read and approved the final manuscript.

Funding This work was funded by the GENES Intra-Africa Academic Mobility Scheme of the European Union (EU-GENES:2017-2552/001-001) program and by a grant from PalmElit SAS.

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

Declarations

Conflict of interest The authors declare no competing interests.

References

- Babu BK, Mathur R (2016) Molecular breeding in oil palm (*Elaeis guineensis*): status and Future perspectives. *Progress Hortic* 48:123–131
- Bai B, Wang L, Zhang YJ et al (2018a) Developing genome-wide SNPs and constructing an ultrahigh-density linkage map in oil palm. *Sci Rep* 8:691. <https://doi.org/10.1038/s41598-017-18613-2>
- Bai B, Zhang YJ, Wang L et al (2018b) Mapping QTL for leaf area in oil palm using genotyping by sequencing. *Tree Genet Genomes* 14:31. <https://doi.org/10.1007/s11295-018-1245-1>
- Ball AD, Stapley J, Dawson DA et al (2010) A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*). *BMC Genomics* 11:1–15
- Ballesta P, Maldonado C, Pérez-Rodríguez P, Mora F (2019) SNP and haplotype-based genomic selection of quantitative traits in *Eucalyptus globulus*. *Plants* 8:331. <https://doi.org/10.3390/plants8090331>
- Barcelos E, SDA R, Cunha RN et al (2015) Oil palm natural diversity and the potential for yield improvement. *Front Plant Sci* 6:190
- Basiron Y (2007) Palm oil production through sustainable plantations. *Eur J Lipid Sci Technol* 109:289–295. <https://doi.org/10.1002/ejlt.200600223>
- Bejarano D, Martínez R, Manrique C et al (2018) Linkage disequilibrium levels and allele frequency distribution in Blanco Orejinegro and Romosinuano Creole cattle using medium density SNP chip data. *Genet Mol Biol* 41:426–433. <https://doi.org/10.1590/1678-4685-GMB-2016-0310>
- Bernardo RN (2010) *Breeding for quantitative traits in plants*, 2nd edn. Stemma Press, Woodbury, Minn
- Billotte N, Jourjon MF, Marseillac N et al (2010) QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis* Jacq.). *TAG Theor Appl Genet Theor Angew Genet* 120:1673–1687. <https://doi.org/10.1007/s00122-010-1284-y>
- Browning BL, Zhou Y, Browning SR (2018) A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet* 103:338–348. <https://doi.org/10.1016/j.ajhg.2018.07.015>
- Caballero A (1994) Developments in the prediction of effective population size. *Heredity* 73:657–679
- Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF (2008) Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178:553–561. <https://doi.org/10.1534/genetics.107.080838>
- Chang L-Y, Toghiani S, Hay EH et al (2019) A weighted genomic relationship matrix based on Fixation Index (FST) prioritized SNPs for genomic selection. *Genes* 10. <https://doi.org/10.3390/genes10110922>
- 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.)
- Cochard B, Adon B, Rekima S et al (2009) Geographic and genetic structure of African oil palm diversity suggests new approaches to breeding. *Tree Genet Genomes* 5:493–504. <https://doi.org/10.1007/s11295-009-0203-3>
- Cochard B, Carrasco-Lacombe C, Pomies V et al (2015) Pedigree-based linkage map in two genetic groups of oil palm. *Tree Genet Genomes* 11. <https://doi.org/10.1007/s11295-015-0893-7>
- Corbin LJ, Liu A, Bishop S, Woolliams J (2012) Estimation of historical effective population size using linkage disequilibria with marker data. *J Anim Breed Genet* 129:257–270
- Corley RHV (2009) How much palm oil do we need? *Environ Sci Policy* 12:134–139. <https://doi.org/10.1016/j.envsci.2008.10.011>
- Corley RHV, Tinker PB (2016) *The oil palm*, 5th edn. Wiley-Blackwell, Chichester
- 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* 18:1–17
- Cros D, Denis M, Sánchez L et al (2015) 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, 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>
- Cuyabano BCD, Su G, Lund MS (2014) Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics* 15:1171. <https://doi.org/10.1186/1471-2164-15-1171>
- Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185:1021–1031. <https://doi.org/10.1534/genetics.110.116855>
- Daval A, Pomiès V, Le Squin S, et al (2021) In silico mapping in an oil palm breeding program reveals a quantitative and complex genetic resistance to *Ganoderma boninense*
- De Roos A, Hayes B, Goddard M (2009) Reliability of genomic predictions across multiple populations. *Genetics* 183:1545–1553
- de Souza LM, dos Santos LHB, Rosa JRBF et al (2018) Linkage disequilibrium and population structure in wild and cultivated populations of rubber tree (*Hevea brasiliensis*). *Front Plant Sci* 9. <https://doi.org/10.3389/fpls.2018.00815>
- Do C, Waples RS, Peel D et al (2014) NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Mol Ecol Resour* 14:209–214. <https://doi.org/10.1111/1755-0998.12157>
- Falconer DS, Mackay TFC (1996) *Introduction to quantitative genetics*, Subsequent edition. Benjamin-Cummings Pub Co, Harlow
- Ferreira A, da SMF, da CE SL, Cruz CD (2006) Estimating the effects of population size and type on the accuracy of genetic maps. *Genet Mol Biol* 29:187–192. <https://doi.org/10.1590/S1415-47572006000100033>
- Flint-Garcia SA, Thornsberry JM, Buckler ES IV (2003) Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* 54:357–374
- Gan ST, Wong WC, Wong CK et al (2018) High density SNP and DaRT-based genetic linkage maps of two closely related oil palm populations. *J Appl Genet* 59:23–34. <https://doi.org/10.1007/s13353-017-0420-7>
- Gerke JP, Edwards JW, Guill KE et al (2015) The genomic impacts of drift and selection for hybrid performance in maize. *Genetics* 201:1201–1211
- Glaubitz JC, Casstevens TM, Lu F et al (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE* 9:e90346. <https://doi.org/10.1371/journal.pone.0090346>
- Gondro C, Werf J van der, Hayes B (eds) (2013) *Genome-wide association studies and genomic prediction*. Humana Press
- 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: volume 1. Managing, sequencing and mining genetic resources*. Springer, Netherlands, Dordrecht, pp 651–682
- Green P, Falls K, Crooks S (1990) Documentation for CRI-MAP, version 2.4
- Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol Biol* 57:461–485. <https://doi.org/10.1007/s11103-005-0257-z>

- Hartley CWS (1988) The oil palm (*Elaeis guineensis* Jacq.). Longman Scientific & Technical ; Wiley, Harlow, Essex
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 92:433–443. <https://doi.org/10.3168/jds.2008-1646>
- He J, Zhao X, Laroche A et al (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front Plant Sci* 5:484. <https://doi.org/10.3389/fpls.2014.00484>
- Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. *Crop Sci* 49:1–12. <https://doi.org/10.2135/cropsci2008.08.0512>
- Herrero J, Santika B, Herrán A et al (2020) Construction of a high density linkage map in oil palm using SPET markers. *Sci Rep* 10:9998. <https://doi.org/10.1038/s41598-020-67118-y>
- Ithnin M, Din AK (eds) (2020) The oil palm genome. Springer International Publishing
- 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* 13:1–14
- Jakobsson M, Edge MD, Rosenberg NA (2013) The relationship between F_{ST} and the frequency of the most frequent allele. *Genetics* 193:515–528. <https://doi.org/10.1534/genetics.112.144758>
- Jeenor S, Volkaert H (2014) Mapping of quantitative trait loci (QTLs) for oil yield using SSRs and gene-based markers in African oil palm (*Elaeis guineensis* Jacq.). *Tree Genet Genomes* 10:1–14. <https://doi.org/10.1007/s11295-013-0655-3>
- Jin J, Lee M, Bai B et al (2016) Draft genome sequence of an elite Dura palm and whole-genome patterns of DNA variation in oil palm. *DNA Res Int J Rapid Publ Rep Genes Genomes* 23:527–533. <https://doi.org/10.1093/dnares/dsw036>
- 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>
- Labate JA, Lamkey KR, Lee M, Woodman WL (1999) Temporal changes in allele frequencies in two reciprocally selected maize populations. *Theor Appl Genet* 99:1166–1178
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>
- Lebedev VG, Lebedeva TN, Chernodubov AI, Shestibratov KA (2020) Genomic selection for forest tree improvement: methods, achievements and perspectives. *Forests* 11:1190. <https://doi.org/10.3390/f11111190>
- Lee M, Xia JH, Zou Z et al (2015) A consensus linkage map of oil palm and a major QTL for stem height. *Sci Rep* 5:8232. <https://doi.org/10.1038/srep08232>
- Li Y, Kim J-J (2015) Effective population size and signatures of selection using bovine 50K SNP chips in Korean native cattle (Hanwoo). *Evol Bioinforma* 11:EBO–S24359
- Lin Z, Hayes BJ, Daetwyler HD (2014) Genomic selection in crops, trees and forages: a review. *Crop Pasture Sci* 65:1177. <https://doi.org/10.1071/CP13363>
- Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* 12:57–63
- Matias FI, Galli G, Granato ISC, Fritsche-Neto R (2017) Genomic prediction of autogamous and allogamous plants by SNPs and haplotypes. *Crop Sci* 57:2951–2958. <https://doi.org/10.2135/cropsci2017.01.0022>
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Montoya C, Lopes R, Flori A et al (2013) Quantitative trait loci (QTLs) analysis of palm oil fatty acid composition in an interspecific pseudo-backcross from *Elaeis oleifera* (H.B.K.) Cortés and oil palm (*Elaeis guineensis* Jacq.). *Tree Genet Genomes* 9:1207–1225. <https://doi.org/10.1007/s11295-013-0629-5>
- Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? *Ann Bot* 110:1303–1316. <https://doi.org/10.1093/aob/mcs109>
- Niu S, Song Q, Koiwa H et al (2019) Genetic diversity, linkage disequilibrium, and population structure analysis of the tea plant (*Camellia sinensis*) from an origin center, Guizhou plateau, using genome-wide SNPs developed by genotyping-by-sequencing. *BMC Plant Biol* 19:1–12
- Nyouma A, Bell JM, 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* 15:1–16
- Nyouma A, Bell JM, Jacob F et al (2020) Genomic predictions improve clonal selection in oil palm (*Elaeis guineensis* Jacq.) hybrids. *Plant Sci* 299:110547. <https://doi.org/10.1016/j.plantsci.2020.110547>
- Nyouma A, Bell JM, Jacob F et al (2022) Improving the accuracy of genomic predictions in an outcrossing species with hybrid cultivars between heterozygote parents: a case study of oil palm (*Elaeis guineensis* Jacq.). *Mol Genet Genomics* 297:523–533
- Ong A-L, Teh C-K, Kwong Q-B et al (2019) Linkage-based genome assembly improvement of oil palm (*Elaeis guineensis*). *Sci Rep* 9:1–9
- Ong A-L, Teh C-K, Mayes S et al (2020) An improved oil palm genome assembly as a valuable resource for crop improvement and comparative genomics in the Arecoideae subfamily. *Plants* 9:1476
- Ouellette LA, Reid RW, Blanchard SG, Brouwer CR (2018) LinkageMapView—rendering high-resolution linkage and QTL maps. *Bioinformatics* 34:306–307. <https://doi.org/10.1093/bioinformatics/btx576>
- Paterson R, Sariah M, Lima N (2013) How will climate change affect oil palm fungal diseases? *Crop Prot* 46:113–120
- Pirker J, Mosnier A, Kraxner F et al (2016) What are the limits to oil palm expansion? *Glob Environ Change* 40:73–81. <https://doi.org/10.1016/j.gloenvcha.2016.06.007>
- 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. <https://doi.org/10.1016/j.ygeno.2015.02.002>
- Porto-Neto LR, Lee SH, Lee HK, Gondro C (2013) Detection of signatures of selection using F_{ST} . In: *Genome-wide association studies and genomic prediction*. Springer, pp 423–436
- Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
- Rajesh MK, Gangurde SS, Pandey MK et al (2021) Insights on genetic diversity, population structure, and linkage disequilibrium in globally diverse coconut accessions using genotyping-by-sequencing. *Omics J Integr Biol* 25:796–809
- Rastas P (2017) Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinforma Oxf Engl* 33:3726–3732. <https://doi.org/10.1093/bioinformatics/btx494>
- Semagn K, Bjørnstad Å, Ndjioudjop MN (2006) Principles, requirements and prospects of genetic mapping in plants. *Afr J Biotechnol* 5. <https://doi.org/10.4314/ajb.v5i25.56082>
- Seng T-Y, Ritter E, Mohamed Saad SH et al (2016) QTLs for oil yield components in an elite oil palm (*Elaeis guineensis*) cross. *Euphytica* 212:399–425. <https://doi.org/10.1007/s10681-016-1771-6>

- Seyum EG, Bille NH, Abteu WG et al (2021) Genome mapping to enhance efficient marker-assisted selection and breeding of the oil palm (*Elaeis guineensis* Jacq.). *Adv Biosci Biotechnol* 12:407–425. <https://doi.org/10.4236/abb.2021.1212026>
- Siberchicot A, Bessy A, Guéguen L, Marais GA (2017) MareyMap Online: a user-friendly web application and database service for estimating recombination rates using physical and genetic maps. *Genome Biol Evol* 9:2506–2509. <https://doi.org/10.1093/gbe/evx178>
- Singh R, Ong-Abdullah M, Low E-TL et al (2013) Oil palm genome sequence reveals divergence of interfertile species in Old and New Worlds. *Nature* 500:335–339. <https://doi.org/10.1038/nature12309>
- Singh R, Tan SG, Panandam JM et al (2009) Mapping quantitative trait loci (QTLs) for fatty acid composition in an interspecific cross of oil palm. *BMC Plant Biol* 9:114. <https://doi.org/10.1186/1471-2229-9-114>
- Slatkin M (2008) Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9:477–485. <https://doi.org/10.1038/nrg2361>
- Soh AC, Mayes S, Roberts JA (eds) (2017) Oil palm breeding: genetics and genomics, 1 edition. CRC Press, Boca Raton
- Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and densities. *J Anim Sci* 86:2447–2454. <https://doi.org/10.2527/jas.2007-0010>
- Sorkheh K, Malysheva-Otto LV, Wirthensohn MG et al (2008) Linkage disequilibrium, genetic association mapping and gene localization in crop plants. *Genet Mol Biol* 31:805–814. <https://doi.org/10.1590/S1415-47572008005000005>
- Statista (2021) Vegetable oils: production worldwide by type, 2012/13-2020/21
- Technow F, Riedelsheimer C, Schrag TA, Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor Appl Genet* 125:1181–1194
- Technow F, Schrag TA, Schipprack W et al (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* 197:1343–1355. <https://doi.org/10.1534/genetics.114.165860>
- Teh C-K, Ong A-L, Kwong Q-B et al (2016) Genome-wide association study identifies three key loci for high mesocarp oil content in perennial crop oil palm. *Sci Rep* 6:19075. <https://doi.org/10.1038/srep19075>
- Teh C-K, Ong A-L, Mayes S et al (2020) Major QTLs for trunk height and correlated agronomic traits provide insights into multiple trait integration in oil palm breeding. *Genes* 11:826
- Teissier M, Larroque H, Brito LF et al (2020) Genomic predictions based on haplotypes fitted as pseudo-SNP for milk production and udder type traits and SCS in French dairy goats. *J Dairy Sci* 103:11559–11573. <https://doi.org/10.3168/jds.2020-18662>
- Tenaillon MI, Austerlitz F, Tenaillon O (2008) Apparent mutational hotspots and long distance linkage disequilibrium resulting from a bottleneck. *J Evol Biol* 21:541–550. <https://doi.org/10.1111/j.1420-9101.2007.01490.x>
- Ting N-C, Jansen J, Mayes S et al (2014) High density SNP and SSR-based genetic maps of two independent oil palm hybrids. *BMC Genomics* 15:309. <https://doi.org/10.1186/1471-2164-15-309>
- Ting N-C, Jansen J, Nagappan J et al (2013) Identification of QTLs associated with callogenesis and embryogenesis in oil palm using genetic linkage maps improved with SSR markers. *PLOS ONE* 8:e53076. <https://doi.org/10.1371/journal.pone.0053076>
- 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. <https://doi.org/10.1186/s12864-015-1985-3>
- Tisné S, Pomiès V, Riou V et al (2017) Identification of ganoderma disease resistance loci using natural field infection of an oil palm multiparental population. *G3 GenesGenomesGenetics* 7:1683–1692. <https://doi.org/10.1534/g3.117.041764>
- Ukoskit K, Chanroj V, Bhusudsawang G et al (2014) Oil palm (*Elaeis guineensis* Jacq.) linkage map, and quantitative trait locus analysis for sex ratio and related traits. *Mol Breed* 33:415–424. <https://doi.org/10.1007/s11032-013-9959-0>
- Wand M (1995) KernSmooth: functions for kernel smoothing supporting. Wand Jones R Package Version 2:23–20
- Waples RS, Do C (2008) Idne: a program for estimating effective population size from data on linkage disequilibrium. *Mol Ecol Resour* 8:753–756. <https://doi.org/10.1111/j.1755-0998.2007.02061.x>
- Watson K, Mayes S, Price Z et al (2001) Quantitative trait loci for yield components in oil palm (*Elaeis guineensis* Jacq.). *Theor Appl Genet* 103:1302–1310. <https://doi.org/10.1007/s122-001-8204-z>
- Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics* 35:235–254. <https://doi.org/10.2307/2529947>
- Weir BS, Goudet J (2017) A unified characterization of population structure and relatedness. *Genetics* 206:2085–2103
- Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193:621–631. <https://doi.org/10.1534/genetics.112.146290>
- Wright S (1931) Evolution in Mendelian populations. *Genetics* 16:97–159
- Xu H, Guan Y (2014) Detecting local haplotype sharing and haplotype association. *Genetics* 197:823–838. <https://doi.org/10.1534/genetics.114.164814>
- Yadav S, Ross EM, Aitken KS et al (2021) A linkage disequilibrium-based approach to position unmapped SNPs in crop species. *BMC Genomics* 22:1–9
- Yamamoto E, Matsunaga H, Onogi A et al (2016) A simulation-based breeding design that uses whole-genome prediction in tomato. *Sci Rep* 6:1–11
- Yan L, Hofmann N, Li S et al (2017) Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC Genomics* 18:529. <https://doi.org/10.1186/s12864-017-3922-0>
- Zheng X, Levine D, Shen J et al (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28:3326–3328. <https://doi.org/10.1093/bioinformatics/bts606>

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



Genomic selection in tropical perennial crops and plantation trees: a review

Essubalew Getachew Seyum · Ngalle Hermine Bille ·
Wosene Gebreselassie Abteu · Norman Munyengwa · Joseph Martin Bell ·
David Cros 

Received: 11 March 2022 / Accepted: 6 September 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract To overcome the multiple challenges currently faced by agriculture, such as climate change and soil deterioration, more efficient plant breeding strategies are required. Genomic selection (GS) is crucial for the genetic improvement of quantitative traits, as it can increase selection intensity, shorten the generation interval, and improve selection accuracy for traits that are difficult to phenotype. Tropical perennial crops and plantation trees are of major economic importance and have consequently been the

subject of many GS articles. In this review, we discuss the factors that affect GS accuracy (statistical models, linkage disequilibrium, information concerning markers, relatedness between training and target populations, the size of the training population, and trait heritability) and the genetic gain expected in these species. The impact of GS will be particularly strong in tropical perennial crops and plantation trees as they have long breeding cycles and constrained selection intensity. Future GS prospects are also discussed. High-throughput phenotyping will allow constructing of large training populations and implementing of phenomic selection. Optimized modeling is needed for longitudinal traits and multi-environment trials. The use of multi-omics, haploblocks, and structural variants will enable going beyond single-locus genotype data. Innovative statistical approaches, like artificial neural networks, are expected to efficiently handle the increasing amounts of heterogeneous multi-scale data. Targeted recombinations on sites identified from profiles of marker effects have the potential to further increase genetic gain. GS can also aid re-domestication and introgression breeding. Finally, GS consortia will play an important role in making the best of these opportunities.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11032-022-01326-4>.

E. G. Seyum · N. H. Bille · J. M. Bell
Department of Plant Biology and Physiology, Faculty of Sciences, University of Yaoundé I, Yaoundé, Cameroon

E. G. Seyum · W. G. Abteu
Department of Horticulture and Plant Sciences, College of Agriculture and Veterinary Medicine, Jimma University, P.O. Box 307, Jimma, Ethiopia

N. Munyengwa
Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane, QLD 4072, Australia

D. Cros (✉)
CIRAD, UMR AGAP Institut, 34398 Montpellier, France
e-mail: david.cros@cirad.fr

D. Cros
UMR AGAP Institut, CIRAD, INRAE, Univ. Montpellier, Institut Agro, 34398 Montpellier, France

Keywords Genomic predictions · Machine learning · Pangenomes · Genotype-by-environment interaction · Crop growth models · Reaction norms

Abbreviations

BLUP	Best linear unbiased prediction
CGM	Crop growth model
CNV	Copy number variation
GBLUP	Genomic BLUP
GEBV	Genomic estimated breeding value
GEGV	Genomic estimated genetic value
GEI	Genotype-by-environment interactions
GS	Genomic selection
GWAS	Genome-wide association study
HTP	High-throughput phenotyping
LD	Linkage disequilibrium
MAS	Marker-assisted selection
NIRS	Near-infrared spectroscopy
NGS	Next-generation sequencing
QTL	Quantitative trait locus
RKHS	Reproducing kernel Hilbert spaces
rrBLUP	Random regression BLUP
SNP	Single nucleotide polymorphism
SV	Structural variants

Introduction

The steady growth of the world population, expected to reach 9–11 billion by 2050, along with climate change and soil deterioration, are major challenges to achieving world food security (Kopittke et al. 2019; Rööß et al. 2017). Biotic and abiotic stresses caused by pathogens, animals, weeds, drought, extreme temperatures, flooding, salinity, acidic conditions, and nutrient starvation all reduce global agricultural productivity (Tyczewska et al. 2018). Plant breeding represents one of the main ways to alleviate these problems and improve both crop production and productivity (Bhat et al. 2016). Plant breeding uses two main approaches, conventional and molecular breeding. Conventional breeding mainly uses phenotypic data (Borrelli et al. 2015) and has several limitations, including the long time (> 10 years) needed to release a new variety, confounding environmental effects leading to low heritability for many traits of interest, particularly the most complex ones, like yield. Molecular plant breeding using DNA markers includes quantitative trait loci (QTL)-based marker-assisted selection (MAS) that can greatly increase the speed, efficiency, and precision of breeding compared to conventional methods (Gupta et al. 2010). However, QTL-based MAS is efficient only for traits

controlled by a few QTLs that have a major effect on trait expression, whereas for complex quantitative traits governed by a large number of minor QTLs, such as yield, it may be less efficient than conventional phenotypic selection (Bhat et al. 2016). For complex traits, the most efficient molecular breeding strategy available today is genomic selection (GS) (Hickey et al. 2019). GS is a form of MAS in which genetic markers covering the whole genome are used so that all QTL are in linkage disequilibrium (LD) with at least one marker (Goddard and Hayes 2007; Heffner et al. 2009; Isik 2014; Meuwissen et al. 2001). GS has emerged as one of the most promising selection strategies to enhance genetic gain per unit time and/or unit cost for both plant and animal breeding programs (Fugerey-Scarbel et al. 2021; Merrick et al. 2022; Mrode et al. 2019; Voss-Fels et al. 2019; Wartha and Lorenz 2021; Xu et al. 2020). In dairy cattle, GS doubled the rate of genetic progress (Wiggans et al. 2017). In plants, GS is progressively integrated into breeding schemes and is now routinely used for major crops, in particular in the private sector (Merrick et al. 2022; Varshney et al. 2017; Voss-Fels et al. 2019). For instance, GS played a key role in the development of drought-tolerant maize hybrids that gave higher yields under both favorable and water stress conditions in the western US Corn Belt (Merrick et al. 2022; Voss-Fels et al. 2019). GS has also been applied on a large scale at the International Maize and Wheat Improvement Center since 2010, where it is used in spring wheat to discard low-performing lines (Merrick et al. 2022).

The first step in GS is creating a training set (or training population). The training set is genotyped and phenotyped for the targeted traits, and a prediction model is then built using these genotypic and phenotypic data. Several high-throughput next-generation sequencing (NGS) technologies such as SNP arrays (LaFramboise 2009; Wang et al. 1998), genotyping-by-sequencing (Elshire et al. 2011), and whole-genome sequencing (Ni et al. 2017) platforms have facilitated the production of large amounts of single nucleotide polymorphism (SNPs) markers to use in GS, at an affordable cost. The target population is also genotyped but not phenotyped, and the prediction model calculates the genomic estimated breeding values (GEBVs) or, when non-additive effects are taken into account, the total genomic estimated genotypic values (GEGV) of the selection candidates

(Grattapaglia et al. 2018). The efficiency of GS is determined, in particular, by its accuracy, which is defined as the correlation between the predicted and the true (unknown) genetic value of the selection candidates (Lorenz et al. 2011). GS accuracy is affected by the effective size of the population, marker density and type, the size and structure of the training population, the genetic architecture of the traits, relatedness between the training and target population, LD between markers and QTLs, trait heritability, imputation method, etc. (Grattapaglia and Resende 2011; Isik 2014; Robertsen et al. 2019).

Tropical perennial crops and plantation trees are of huge importance for the human population, in particular for use as food, timber, pulp, and stimulant crops (Jamnadass et al. 2016). However, their productivity is generally well below their potential, in particular, due to biotic and abiotic constraints, as shown, for example, in Eucalyptus (Elli et al. 2019), oil palm (Pirker et al. 2016; Woittiez et al. 2017), coffee (Wang et al. 2015), and cocoa (Aneani and Ofori-Frimpong 2013). Applying more efficient breeding approaches to these species will help fill production gaps. Genomic selection is particularly attractive for perennial plant species as they have long generation intervals and low selection intensity. Isik (2014) showed that the impact of GS could be much greater in perennial forest trees than in any other crop or livestock breeding program. A significant number of articles on GS have already been published on a variety of traits of interest in several tropical perennial crops and plantation trees, for instance, yield in oil palm (Cros et al. 2017, 2015), rubber tree (Cros et al. 2019) and guava (Silva et al. 2021), growth in eucalyptus (Bouvet et al. 2016; Denis et al. 2012; Resende et al. 2012) and rubber tree (Souza et al. 2019), fruit quality in citrus (Minamikawa et al. 2017), resistance to diseases in cocoa (McElroy et al. 2018; Romero Navarro et al. 2017), etc. (Supplementary Table S1). However, a review of GS in these species is lacking. The objective of the present article is therefore to review the results of GS research in tropical perennial crops and plantation trees, to discuss the main factors affecting GS accuracy and to highlight the genetic gains expected in these species using this approach. We focus on perennial crops defined as such according to the FAO indicative crop classification (FAO 2015) and on plantation trees both grown in the tropics. The production of the corresponding

species include fruit, timber, pulp, latex, oil, nuts, and stimulants. To our knowledge, the species covered by published articles on GS so far are banana, guava, citrus, *Eucalyptus* species (*E. urophylla*, *E. grandis*, *E. benthamii*, *E. pellita*, and *E. robusta*), rubber tree, oil palm, jatropha, cacao, and coffee.

Factors affecting the accuracy of genomic selection

The correlation between the GEBVs and true breeding values is known as GS accuracy (r_{GS}), and it is a key parameter for breeders due to the linear correlation between selection accuracy and annual genetic gain R_y (Eq. (1)) (Grattapaglia et al. 2018):

$$R_y = \frac{i \times r \times \delta_A}{y} \quad (1)$$

where i is selection intensity, r is selection accuracy, δ_A is the additive genetic standard deviation, and y is the generation interval in years.

GS accuracy is usually obtained by k-fold cross-validation within a single experimental design (with each fold repeatedly used as a validation set and the remaining folds as the training set) or between experimental designs (with one site used for training and the other for validation), the latter being preferable as cross-validations may overestimate accuracy (Lorenz et al. 2011).

Below, we present sequentially the major factors that affect the accuracy of genomic predictions, although most factors are interconnected and their effects are not independent.

Statistical models for genomic prediction and trait genetic architecture

The whole-genome regression models used for genomic predictions deal with the “large p , small n ” problem that, in GS, concerns the number of markers that usually (largely) exceeds the number of data records, in contrast to multiple linear regressions that cannot be used without variable selection, which conflicts with the original goal of GS, i.e., avoiding marker selection and overfitting. Multiple linear regression results in an insufficient degree of freedom leading to poor prediction due to the inability to estimate all marker effects at

the same time, which is exacerbated by multicollinearity. A wide range of statistical methods has been developed for GS to alleviate this constraint (Campos et al. 2013; Jannink et al. 2010; Montesinos-López et al. 2021; Morota and Gianola 2014; Tong and Nikoloski 2021; Wang et al. 2018). They represent two broad categories: (i) parametric approaches, which mainly include methods that rely on the best linear unbiased prediction methodology (genomic BLUP [GBLUP] and random regression BLUP [RRBLUP]) and various Bayesian methods (Bayesian LASSO, BayesA, BayesB, etc.), and (ii) semi- and non-parametric approaches that fall into the machine learning category (reproducing kernel Hilbert spaces [RKHS], artificial neural networks, etc.). These methods differ in several ways: in terms of genetic assumptions and modeling of the genetic architecture of the traits (e.g., purely additive models, models that explicitly model dominance and/or epistatic effects, models with marker effects sampled from a common statistical distribution [RRBLUP, GBLUP], models with marker effects sampled from specific distributions [Bayesian LASSO, BayesB, etc.], models that implicitly model non-additive effects [e.g., RKHS]), in terms of computational approach (relationship-based methods and marker effect-based methods, single trait and multi-trait models, etc.), and in terms of the genomic information used in the model (type of polymorphisms, use of a priori information on markers, a combination of omics data, etc.).

The most widely used statistical approach for GS is GBLUP (Heslot et al. 2015; Montesinos-López et al. 2021), which combines linear mixed model analysis and genomic relationships. GBLUP derives from the first BLUP analyses applied in animal breeding to implement selection based on phenotypes and pedigree and that estimated the breeding values of individuals using the pedigree-based relationship matrix (A) (Henderson 1975), with a model of the form:

$$Y = X\beta + Zu + e \quad (2)$$

where Y is an $n \times 1$ vector of data records, X is an $n \times p$ incidence matrix relating data records with fixed effects, β is a $p \times 1$ vector of fixed effects, and Z is an $n \times q$ incidence matrix. u is a $q \times 1$ vector of random effects (i.e., breeding values), associated

with A , and e is an $n \times 1$ vector of residual effects. This initial approach we term pedigree-based BLUP (PBLUP) paved the way for GBLUP, which uses the genomic relationships (G) matrix, thus capturing existing relationships among individuals rather than expected relationships (Bernardo 1994; VanRaden 2007). An alternative approach to GBLUP is RRBLUP (Meuwissen et al. 2001), which yields GEBVs by estimating marker effects. GBLUP and RRBLUP are equivalent when there are many QTLs, when there is no major QTL, or when the QTLs are evenly distributed along the genome (Bernardo 2020). RRBLUP uses a model of the form:

$$Y = X\beta + Z'm + e \quad (3)$$

where Z' is an $n \times k$ incidence matrix giving the genotypes at k SNPs and m a $k \times 1$ vector of random SNP effects.

The relative performance of the different statistical methods is expected to vary depending on the genetic architecture of the trait considered (Lebedev et al. 2020). Genetic architecture corresponds to the genetic characteristics that determine the genotype–phenotype relationship, in particular, the number of genes that control the trait, the number of alleles per gene, the distribution of the genes along the genome, the distribution of the gene effects, and the mode of gene action (additive, dominant, epistatic) (Momen et al. 2018). Thus, methods in which marker effects are sampled in distributions where variance is the same for all markers (e.g., GBLUP, RRBLUP, Bayesian random regression) are expected to be more suitable for traits following the infinitesimal model, while methods with marker-specific variances (e.g., Bayesian LASSO, BayesB) are expected to be more suitable for traits whose genetic architecture includes major QTLs. Consequently, many GS studies, including those on tropical perennial fruit crops and plantation trees, use a range of statistical prediction methods to identify the most appropriate one for a specific trait. Overall, few variations have been found among statistical approaches, for example, in oil palm yield components (Cros et al. 2015; Kwong et al. 2017a), in eucalyptus growth (Durán et al. 2017; Müller et al. 2017), and in rubber tree latex yield (Cros et al. 2019). This confirms results obtained in empirical evaluations in other species, in which GS statistical methods were seen to perform similarly

(Heslot et al. 2015); however, in some cases, differences were found: e.g., BayesB performed best for several traits including vegetative growth, production, and disease resistance in banana (Nyine et al. 2018) and vegetative growth and oil yield in oil palm (Ithnin et al. 2017). This could mean that, in the populations considered, QTLs with large effects were segregated for these traits.

Similarly, when non-additive effects play a significant role in genetic variation, models that account for non-additive effects are expected to increase GS accuracy. In a simulation study, Denis and Bouvet (2013) showed that modeling dominance for the genomic predictions of the genetic value of eucalyptus clones improved accuracy when dominance effects were preeminent (ratio of dominance to the additive variance of 1.0) and heritability was high ($H^2=0.60$). With empirical data, also in eucalyptus, Resende et al. (2017), Tan et al. (2018), and Paludeto et al. (2021) showed that the use of GS models that account for dominance increased the accuracy of prediction for growth traits, which had high levels of dominance variance, whereas this was not the case for wood traits. In citrus, Minamikawa et al. (2017) showed that considering both additive and dominance effects improved prediction accuracy for acidity and juiciness.

When considering traits correlated with a sufficient magnitude but with contrasting levels of heritability, the use of multi-trait models can increase prediction accuracy for low heritability traits (Tong and Nikoloski 2021). In tropical perennial crops and plantation trees, the results obtained in oil palm (Marchal et al. 2016) and *Eucalyptus robusta* (Rambolarimanana et al. 2018) agreed with this principle. Multivariate models thus offer the opportunity to improve prediction accuracy at no extra cost (apart from increased computational resources), and they should therefore be systematically evaluated when correlations exist among the traits of interest, or between the traits of interest and secondary traits.

Machine learning methods are complex black-box approaches that are of growing interest for genomic predictions as they have several desirable features. They avoid the use of assumptions that are often violated and cannot be verified (Gianola and Van Kaam 2008), and they are particularly suitable to account for non-additive effects in particular in polyploids (Bayer et al. 2021) and to integrate data

from different biological sources for multi-omics predictions (Montesinos-López et al. 2021; Tong and Nikoloski 2021). RKHS is the most often evaluated machine learning approach for GS in tropical perennial crops and plantation trees. In bananas, RKHS was slightly more accurate than parametric approaches for a few traits (Nyine et al. 2018). In a study analyzing eight traits in *E. urophylla* × *E. grandis* eucalyptus hybrids, RKHS proved to be slightly more accurate in predicting low-heritability traits but less accurate in predicting pulp yield (Tan et al. 2017) and performed similarly to GBLUP for three traits in *E. grandis* (Rambolarimanana et al. 2018). A few other machine learning methods have been implemented in tropical perennial crops and plantation trees. Maldonado et al. (2020) compared several parametric prediction models, RKHS and two artificial neural network approaches, deep learning and Bayesian regularized neural networks, in *E. globulus* and maize, and found that predictions made with deep learning methods were significantly more accurate for all the traits considered. Sousa et al. (2020) compared several machine learning approaches and a parametric model to predict resistance to leaf rust in *Coffea arabica* and obtained the best accuracy with artificial neural networks. Several authors used random forest in oil palm and citrus and found that, on average over several traits, random forest performed no better than parametric approaches (Kwong et al. 2017b; Minamikawa et al. 2017). In oil palm, the support vector machine was found to be slightly better on average than other methods (Kwong et al. 2017b). Despite these uneven results in tropical perennial crops and plantation trees, machine learning should be further investigated, in particular as the training populations used so far were possibly not large enough for the optimal training of this type of approach (Montesinos-López et al. 2021). Particular attention should also be paid to artificial neural networks, which have produced promising results.

One limit to the differences among statistical methods and models in perennial fruit and tree crops reported so far is that they were not always supported by a statistical test indicating whether the differences were significant or not. This can be done, for example, using the Hotelling-Williams t -test (Steiger 1980).

Linkage disequilibrium and effective size

Linkage disequilibrium (LD) between markers and QTL and effective size (N_e) have interrelated effects that strongly influence GS accuracy (Heffner et al. 2009; Isik 2014; Lebedev et al. 2020). LD is defined as the non-random association of alleles at two or more loci in haplotypes (Slatkin 2008; Weir 1979). LD between two loci is measured based on the frequency of alleles, using indexes like D , D' , and r^2 (Collins and (Ed.) 2007). A key assumption in GS is that there is LD between QTLs and markers, such that, with dense genome marker coverage, every QTL controlling the phenotype of interest would be in LD with at least one marker. Good knowledge of this parameter in the target population is therefore of particular interest to define the marker density required for GS. It is thus useful to explore historical events, such as bottlenecks, genetic drift, and natural and artificial selection, that may have shaped the LD profile in the target population (Flint-Garcia et al. 2003; Gupta et al. 2005; Mackay and Powell 2007; Slatkin 2008). The LD profile is largely determined by the past N_e , which can be described as the number of randomly mating individuals in a population that would give rise to the observed rate of inbreeding (Falconer and Mackay 1996). There is an inverse relationship between N_e and LD, with high rates of genetic drift and inbreeding in low N_e populations leading to strong LD between markers and QTLs compared to high N_e populations (Grattapaglia 2014; Lin et al. 2014; Thistlethwaite et al. 2020). As N_e decreases and LD increases, pairs of individuals within the population tend to share longer haplotypes, enabling good genomic prediction accuracy (Clark et al. 2012; Heffner et al. 2009; Isik 2014; Lebedev et al. 2020). For a given marker density, training population size, and trait, LD and GS prediction accuracy is higher in populations with low N_e than in populations with high N_e (Grattapaglia 2014; Lin et al. 2014; Solberg et al. 2008).

The crucial role of LD and N_e in GS accuracy has also been underlined in studies on tropical perennial crops and plantation trees. Several studies investigated the LD profile to evaluate whether the marker density was high enough in citrus (Gois et al. 2016; Minamikawa et al. 2017), cocoa (McElroy et al. 2018), eucalyptus (Denis and Bouvet 2013; Durán et al. 2017; Müller et al. 2017), and oil palm

(Kwong et al. 2017a). Many studies in tropical perennial crops and plantation trees also investigated the efficiency of GS in populations with high LD/low N_e . This was possible using populations obtained through specific mating designs among a reduced number of parents (Denis and Bouvet 2013; Resende et al. 2012). In this way, Resende et al. (2012) found that in a population of eucalyptus where $N_e=11$ was obtained with an incomplete diallel, GS accuracy was higher for the four growth and wood quality traits studied than in the population where $N_e=51$, despite a slightly larger number of training individuals in the latter population. In other studies, high LD/low N_e was obtained in full-sib families GS (Cros et al. 2017; de Souza et al. 2018; Gois et al. 2016; Kwong et al. 2017b). This strategy is also applied in other crops as it maximizes GS accuracy, although at the cost of only applying to families comprising the training population (Crossa et al. 2017; Lebedev et al. 2020; Lin et al. 2014).

The fact that GS accuracy reaches a plateau when marker density reaches a certain level (see below) suggests that an appropriate strategy to filter the markers would increase the cost-efficiency of GS. Filtering SNPs on LD has been investigated in several studies, as the SNPs that show very high LD values provide redundant information. In oil palm, Kwong et al. (2017a) evaluated the impact of marker density reduction by LD filtering and noted that, for some traits, it was possible to reach the same GS accuracy as using all the SNPs.

Marker density and marker type

As marker density strongly affects the extent of LD, it also plays a major role in GS accuracy. In GS studies of both plants and animals, increasing the number of markers was shown to improve prediction accuracy until a plateau was reached (Isik 2014; Lin et al. 2014; Meuwissen et al. 2001; Robertsen et al. 2019; Solberg et al. 2008). The same trend was observed in tropical perennial crops and plantation trees, where the density of markers required to reach maximum prediction accuracy depends in particular on the type of population, trait, and marker. Romero Navarro et al. (2017) found increasing prediction accuracy for yield and disease traits in cocoa with increasing marker density before a plateau was reached at around 1000 markers. In the rubber tree,

the prediction accuracy for rubber yield plateaued at around 300 SRRs (Cros et al. 2019). In eucalyptus, the prediction accuracy among five growth and wood property traits reached a plateau between 5000 and 20,000 SNPs (Tan et al. 2017). Among seven production traits in oil palm hybrids, the plateau was reached with 500 to 2000 SNPs (Cros et al. 2017).

GS accuracy is also affected by the type of marker. Thus, in oil palm, GS accuracy for bunch number and average bunch weight plateaued at 160 SSRs in heterotic group A and at 90 SSRs in group B (Marchal et al. 2016) versus 3000 SNPs in group A and 350 SNPs in group B (Cros et al. 2017). This likely resulted from the fact that, as SNPs are biallelic, they are less informative than SSRs. However, in practice, SSRs cannot be used for genomic predictions, as GS relies on dense genotyping of large populations of selection candidates and therefore requires high throughput genotyping approaches at a reasonable cost. If marker density is constrained by the genotyping approach, the GS accuracy may be reduced. Thus, Kwong et al. (2017b) obtained mean GS prediction accuracies of 0.21 over palm oil yield components using 135 SSRs, versus 0.31 with 200 K SNPs.

Two primary options are available to reach the high marker density required for GS: methods that reduce genome complexity and SNP arrays (Edwards et al. 2013; Wiggans et al. 2017). They were made possible by the development of NGS technologies, which became available between 2004 and 2006 (Hu et al. 2021). Less expensive and with much higher throughput than the Sanger method (Sanger and Coulson 1975; Sanger et al. 1977), NGS methods have made it possible to carry out high-density and high-throughput genotyping, i.e., with good genome coverage in large populations, at an affordable cost. SNP arrays have been developed in several tropical perennial crops and plantation trees, with, for example, a 200 K array in oil palm (Kwong et al. 2016), a 60 K array in eucalyptus (Silva-Junior et al. 2015), and a 15 K array in cacao (McElroy et al. 2018). Most SNP genotyping methods based on reducing genome complexity consist of restriction enzyme-based approaches and sequence capture (Uitdewilligen et al. 2013; Zhou and Holliday 2012). These methods do not require specific preliminary investment and can be applied directly to any population. Given their relative simplicity and lower cost compared to SNP

arrays, they became widely used, in particular for introgression breeding, genome-wide association mapping (GWAS), and QTL mapping (see, e.g., Kitony et al. (2021) and Reyes et al. (2021) in rice, Pootakham et al. (2015) in oil palm, or Chia Wong et al. (2022) in cacao). However, they are associated with a higher rate of missing data and genotyping errors than SNP arrays. Despite these differences, it seems that the choice between these two types of approaches has no impact on GS accuracy: The accuracy of genomic prediction of 13 wood quality and growth traits in eucalyptus using SNP genotypes obtained with sequence capture and a 60 K SNP array was similar (de Moraes et al. 2018).

Training and validation population relatedness

The accuracy of GS is positively correlated with the relatedness between the training and test population (Daetwyler et al. 2013; Isidro y Sánchez J, Akdemir D 2021; Pszczola et al. 2012; Wientjes et al. 2013). This is because when pairs of genotypes are closely related, they tend to share long haplotype blocks in the same linkage phase. To limit allele duplication and redundancy, relationships within the training population should be minimized (Isidro y Sánchez J, Akdemir D 2021). The accuracy of GS in tropical perennial crops and plantation trees was also found to be affected by the relatedness between the training and test population. In two eucalyptus species, *E. benthamii* and *E. pellita*, Müller et al. (2017) found that prediction accuracy declined strongly for three growth traits when individuals were randomly assigned to the training and validation populations compared to when they were assigned using a principal component analysis to minimize relatedness between training and validation populations. Similarly, considering eight wood growth and quality traits in *Eucalyptus urophylla* × *E. grandis*, Tan et al. (2017) obtained the worst prediction accuracies when minimizing the relatedness between the training and validation populations using k-means clustering. In another study, a significant positive correlation was found between GS accuracy and the relationship between training and validation populations for various production traits in oil palm (Cros et al. 2015).

Size and design of the training population

The size of the training population is one of the most important factors that determine GS accuracy. Several GS studies have reported that increasing the size of the training population improves GS accuracy (Calleja-Rodriguez et al. 2020; Cericola et al. 2018; Combs and Bernardo 2013; Isidro et al. 2015; Liu et al. 2018; Nielsen et al. 2016; Tan et al. 2017). In a family of full-sibs of *Hevea brasiliensis*, Cros et al. (2019) reported an increase in the accuracy of GS for rubber yield with an increase in the size of the training population up to a plateau of 200 individuals. In Eucalyptus, Denis and Bouvet (2013) also reported an increase in GS accuracy as a result of increasing the size of the training population, and Tan et al. (2017) reported an increase in GS accuracy that followed a diminishing return trend with increasing size of the training population.

The possibility of assembling large training populations among tropical perennial crops and plantation trees is contrasted. Thus, training populations comprising more than 1000 individuals were used in eucalyptus (Mphahlele et al. 2021), cacao (McElroy et al. 2018), and oil palm (Kwong et al. 2017a), whereas only small populations (<600 individuals) have been used so far in banana (Nyine et al. 2018), rubber tree (Cros et al. 2019; Munyengwa et al. 2021; Souza et al. 2019), coffee (Fanelli Carvalho et al. 2020; Ferrão et al. 2019; Sousa et al. 2020, 2019, p. 2), jatropha (Peixoto et al. 2017), and guava (Silva et al. 2021). However, the size of the training population must be considered in relation to the relatedness between training and validation populations. Thus, for GS predictions in a biparental cross, it is better to use a relatively small but highly related training population of full-sibs or half-sibs than a large training population comprising distantly related or unrelated individuals (Brandariz and Bernardo 2019a; Brauner et al. 2020).

For some of the species considered here, breeding relies on a large number of phenotyped individuals, e.g., thousands of individuals for yield components and tolerance to ganoderma disease in oil palm (Cros et al. 2017; Daval et al. 2021) and thousands of individuals for tolerance to pests and diseases in *Eucalyptus grandis* (Mphahlele et al. 2021). In this case, genotyping a sample of the phenotyped population and making the genomic predictions using the single-step

GBLUP approach (Lourenco et al. 2020), i.e., using a training population combining the genomic data of the genotyped individuals and the genealogical data of the others, is an efficient way to maximize the cost-efficiency of GS; see Mphahlele et al. (2021) in *E. grandis*, Cappa et al. (2019) in a complex eucalyptus population, and Imai et al. (2019) in citrus.

The cost of phenotyping is a major constraint in GS, especially now that sequencing costs have dramatically decreased thanks to next-generation sequencing (Akdemir and Isidro-Sánchez 2019). This financial constraint is particularly applicable to perennial crops, as their phenotypic evaluation requires large surface areas over several years. Thus, training populations need to be optimized to improve the cost-effectiveness of GS in these species. Training population optimization is the process of selecting, within a pool of individuals that could be used to train the GS model, a sample of individuals that will best predict the genetic value of the selection candidates (Isidro y Sánchez J, Akdemir D 2021). Several methods have been developed to optimize the training population, including CD-mean, PEV-mean, stratified sampling, or EthAcc (Isidro y Sánchez J, Akdemir D 2021). This aspect has received little attention in tropical perennial crops and plantation trees, although in oil palm, Cros et al. 2015 confirmed the efficiency of training population optimization to improve GS accuracy.

Trait heritability

The broad-sense heritability of a trait (H^2) is defined as the proportion of the phenotypic variance that is genetically controlled. Narrow-sense heritability (h^2) considers only variations due to additive gene action and ignores non-additive (dominance and epistasis) genetic effects (Falconer and Mackay 1996). In GS studies, the heritability of the trait affects the accuracy of GEBV, with higher h^2 leading to greater GS accuracy (Hayes et al. 2009; Lin et al. 2014; Meuwissen et al. 2001, p. 2). This was illustrated by studies in tropical perennial crops and plantation trees where positive correlations were found between h^2 and GS prediction accuracy for a set of disease resistance and yield traits in cacao (Romero Navarro et al. 2017), eight palm oil production traits in the B heterotic group used in oil palm breeding (Cros et al. 2015), 18 Arabica coffee agronomic traits (Sousa et al., 2019),

and 15 vegetative growth, disease resistance, and fruit production traits in banana (Nyine et al. 2018). When simulating GS in eucalyptus, Denis and Bouvet (2013) noted that the prediction accuracy was higher with $H^2=0.6$ than with $H^2=0.1$, regardless of the ratio of dominance to additive variance, modeling dominance or not, or the breeding cycle. However, some studies detected no effect of trait heritability on GS prediction accuracy, but the effect may have been masked by other factors with stronger effects on prediction accuracy than heritability, in particular variations in the size of the training population, among traits, like in Durán et al. (2017).

Genetic gain from genomic selection

Genetic gain from the selection is defined as the improvement in the average genetic value of a population under the effect of selection over breeding cycles (Hazel and Lush 1942). GS has substantially increased genetic gain in animal breeding and plays a central role in many commercial plant breeding programs (Fugerey-Scarbel et al. 2021; Voss-Fels et al. 2019; Wartha and Lorenz 2021; Xu et al. 2020). The main advantages of GS over conventional phenotypic selection are its ability to (i) increase selection intensity and/or to shorten the generation interval by replacing all or part of the phenotyping activities by genotyping in selected breeding cycles and (ii) increase accuracy for traits that are difficult to phenotype (Fugerey-Scarbel et al. 2021; Wartha and Lorenz 2021).

When GS is used to increase selection intensity or to shorten the breeding cycle, an increase in annual genetic gain can be obtained even though GS is less accurate than conventional phenotypic evaluation. This has been illustrated in studies of tropical perennial crops and plantation trees that are promising for GS due to their long generation intervals and challenging phenotypic evaluations. Thus, based on the relative accuracy of GS and phenotypic selection, Resende et al. (2012, 2017) demonstrated that GS could significantly increase annual genetic gain for growth and wood quality traits in eucalyptus, i.e., from +50% to +300%, thanks to the fact that GS can be implemented at the seedling stage (<1 year), i.e., much earlier than phenotypic selection, which cannot be carried out before at least three years old.

Additionally, the possibility of increasing selection intensity by using a bigger population of selection candidates should further increase the advantage of GS over conventional selection. Based on 17 years of *E. grandis* breeding, Mphahlele et al. (2021) reported that the accumulated genetic gain with GS would be from 1.53 to 3.35 times higher than with conventional phenotypic selection, depending on the trait, because GS allows three breeding cycles in a 17-year period versus two with phenotypic selection. In coffee, it was also shown that with GS, 3-year breeding cycles would lead to a higher annual genetic gain in traits for growth, production, and tolerance to biotic stresses than the conventional 6-year phenotypic breeding cycles in *Coffea arabica* (Sousa et al. 2019) and in *Coffea canephora* (Alkimim et al. 2020). Similarly, an increase in annual genetic gain through a reduction in the generation interval with GS has been reported in citrus (Gois et al. 2016) and in rubber tree (Souza et al. 2019).

However, in many cases, the advantage of using GS over phenotypic selection in terms of genetic gain did not concern all the traits of interest. In this case, the interest of GS is its ability to increase selection intensity. This leads to a two-stage breeding scheme, starting with genomic selection, followed by phenotypic selection. In this case, the limiting factor for GS is the number of selection candidates that can be genotyped. In oil palm, using GS for bunch production before conventional phenotypic progeny tests was estimated to improve the performance of the selected A×B hybrids by more than 10% when 4000 A and 4000 B were genotyped (Cros et al. 2017). Similarly, in a full-sib rubber tree family, applying GS to 3000 individuals before clonal trials would have increased the selection response for rubber production by around 10% (Cros et al. 2019).

Some studies on tropical perennial crops and plantation trees also compared GS and QTL-based MAS approaches and the genetic gain expected from GS. For instance, in cacao, McElroy et al. (2018) found that GS largely outperformed GWAS in genetic gain for most of the disease resistance traits considered. In breeding populations of eucalyptus under selection, Müller et al. (2017) showed that GS outperformed GWAS for growth traits, as GS accounted for large proportions of the heritability, whereas GWAS captured very few significant associations. In a study simulating several cycles of within-family oil palm

breeding, Wong and Bernardo (2008) found that GS enabled higher annual genetic gains than marker-assisted recurrent selection for all the family sizes, number of QTLs, and heritability considered.

Future prospects for genomic selection in perennial tropical crops and plantation trees

Promising results have already been obtained with GS in tropical perennial crops and plantation trees. However, different aspects require further investigation to take full advantage of the approach. As mentioned above, statistical approaches for predictions still require attention; in particular, single-step GBLUP and multivariate models need to be more widely used and artificial neural networks need to be investigated in greater detail. Training populations also need optimization. Other promising aspects have hardly or not been studied at all so far for use with GS in tropical perennial crops and plantation trees, and these aspects are discussed below.

High-throughput phenotyping

High-throughput phenotyping (HTP) platforms allow faster phenotyping and reduced labor costs compared to conventional methods (Persa et al. 2021). HTP allows analyses at the field scale with outdoor platforms that use remote sensing and imaging, mostly based on visible/near-infrared and far-infrared spectroscopy, and analyses of the harvestable part of the crop using near-infrared reflectance spectroscopy (NIRS). The use of HTP has already led to significant results in model species such as rice, maize, and wheat, for a wide range of traits, like adaptation, quality, and vegetative growth (Asaari et al. 2019; Blancon et al. 2019; Chattopadhyay et al. 2019; Juliana et al. 2019; Sun et al. 2019; Wu et al. 2019). For GS, HTP is an efficient way to characterize large training populations (Wartha and Lorenz 2021). This is particularly useful for perennial species that require phenotyping over extended periods of time. HTP has already been used in different tropical perennial crops and plantation trees. For instance, multispectral data collected from an unmanned aerial vehicle were used to estimate the height and diameter at the breast height of eucalyptus trees (Borges et al. 2021). NIRS has also been used for rapid quantification of

flavor-related components of cocoa and beverage quality components of Arabica coffee (e.g., Álvarez et al. 2012; dos Santos Scholz et al. 2014). In eucalyptus populations used for GS, NIRS was used to measure chemical and physical wood quality traits (de Moraes et al. 2018; Durán et al. 2017; Rambolari-manana et al. 2018).

In addition to enabling the phenotyping of large populations, HTP data can be used in GS models as covariates associated with the trait of interest to increase prediction accuracy (Persa et al., 2021). To our knowledge, this aspect has not been investigated so far in GS studies on tropical perennial crops and plantation trees, but such studies would be of interest.

Phenomic selection is another approach that relies on spectral data that are usually obtained by NIRS (Rincint et al. 2018). In this case, the prediction of the genetic values is based on spectral data instead of molecular markers, meaning genomic data could no longer be needed. Phenomic selection has been investigated in a few crops, particularly in two temperate perennial species, poplar and grapevine. In poplar, the expected genetic gain using phenomic selection was higher than or the same as using genomic selection, depending on the trait (Rincint et al. 2018). In grapevine, phenomic predictions were reported to be a possible alternative to genomic predictions (Brault et al. 2022).

Longitudinal traits

Longitudinal traits are traits recorded repeatedly over the period of interest in the lifetime of individuals. This is a common case in perennial species. In tropical perennial crops and plantation trees, longitudinal traits are, for instance, growth and production, which are evaluated on each plant at different ages. The random regression model, a standard approach used for the genetic analysis of such traits (Oliveira et al. 2019), is a mixed model that makes it possible to model individual genetic values as a continuous function of time (or environmental covariates, see below), which can lead to more accurate estimates of the genetic values and facilitate the selection of genotypes with an optimal profile over the period of interest. Random regression can link genetic effects and time with complex functions, including nonlinear patterns, without making assumptions about the shape of the curve (Mrode 2014; Oliveira et al. 2019).

The parameters that characterize these functions (e.g., slopes and intercepts for linear functions) are treated as random effects, and the analysis yields genotype-specific parameters. Random regression has already been used for genomic predictions of longitudinal traits in different species, in particular in animals (Oliveira et al. 2019). Surprisingly, even though many traits in tropical perennial crops and plantation trees are longitudinal, random regression has rarely been used in these species. One example is *Jatropha curcas*, where random regression was used to analyze grain yield over the years (Peixoto et al. 2020). However, to our knowledge, this approach has not been used in the context of GS in tropical perennial crops and plantation trees so far.

Leveraging multi-environment trials

Multi-environment trials and GS models that account for environmental effects make it possible to predict the genetic value of new genotypes in known environments, known genotypes in new environments, and new genotypes in new environments (Bustos-Korts et al. 2016; Malosetti et al. 2016). The ability to predict the performances in new environments is of major interest in the context of climate change, in particular for perennial crops where breeding suffers from inertia due to the length of the breeding cycles. Analysis of genotype-by-environment interactions (GEI) helps select genotypes that are stable across environments and can identify the best genotypes for specific target environments. In particular, this has been extensively studied in cereals (Crossa et al. 2017). Considering GEI in GS models can significantly increase prediction accuracy when data from multi-environment trials are available (Tong and Nikoloski 2021; Xu et al. 2020). A variety of approaches have been developed to incorporate environmental data in GS models (Bustos-Korts et al. 2016; Crossa et al. 2017; Malosetti et al. 2016; Tong and Nikoloski 2021; Xu et al. 2020). The most attractive methods enable predictions in new environments using reaction norms (Costa-Neto et al. 2021; Costa-Neto and Fritsche-Neto 2021; Crossa et al. 2021) or crop growth models (CGM) (Crossa et al. 2021; Van Eeuwijk et al. 2019; Xu et al. 2020).

Reaction norms are linear or nonlinear functions that describe the phenotypes produced by a single genotype across an environmental gradient (Li

et al. 2017). They can be incorporated into genetic analyses using random regression (Marchal et al. 2019; Mrode 2014; Oliveira et al. 2019), leading to genotype-specific coefficients that characterize random norms for each environmental covariate. Equivalently, the environmental covariates can be used to build an environmental relationship matrix that identifies putative similarities among the environments considered (Costa-Neto et al. 2021), rather like using SNPs to build the relationship matrix.

CGM relies on plant physiology, soil science, and climatology principles to model plant development. CGMs use equations involving genetic parameters that are specific to the genotypes under consideration and are assumed to be independent of the environment and environmental variables (Boote et al. 2013). Several methods have been developed to incorporate CGM in the context of GS (Crossa et al. 2021; Rincent et al. 2017). CGM can be implemented to predict developmental stages that – along with daily weather data – will be used to compute climate stress covariates according to the plant development stage. CGM can also be used to compute environmental stress covariates that include the response of the crop to environmental conditions. These environmental covariates can then be incorporated in the GS model using, for example, random regression. Alternatively, the genetic parameters of the CGM can be estimated for the genotypes that comprise the training set and the genetic parameters of the selection candidates predicted by a GS model. Using the CGM and environmental covariates makes it possible to predict the phenotype of the selection candidates in the target environment. This approach has been termed gene-based modeling. Another method consists of incorporating a CGM in the GS prediction framework for the joint estimation of marker effects and CGM genetic parameters. This is referred to as CGM-WGP (whole-genome predictions) and relies on the use of approximate Bayesian computation or Bayesian generalized linear hierarchical models.

Ideally, the use of reaction norms or CGM requires the identification of all the environmental covariates that affect the trait of interest and the availability of environmental data at the plant level. This refers to the concept of envirotyping (Xu 2016) and its extension to large scale across time and space and environments (Resende et al. 2021). To our knowledge, only two GS studies have considered multi-environment

trials in tropical perennial crops and plantation trees so far. Souza et al. (2019) made genomic predictions obtained with multi-environment data and modeling approaches including environmental effects and GEI applied to rubber trees grown in two environmental conditions. These authors showed that multi-environment models captured a larger proportion of the genetic variance than single-environment approaches. In *Coffea canephora*, Ferrão et al. (2019) used multiplicative models in which genetic and environmental effects were handled in a common random effect associated with a variance–covariance matrix obtained by the Kronecker product of genetic and environmental variance–covariance matrices. These authors showed that this approach resulted in more accurate GS than traditional GBLUP, as the latter did not account for environmental information. This area of GS needs further study in tropical perennial crops and plantation trees, and particular attention should be paid to the use of CGM, reactions norms, and enviromics. This could leverage tools and skills that are already

available in these species. Thus, crop growth models have already been developed, for example, in cocoa (Zuidema et al. 2005), oil palm (Huth et al. 2014), and eucalyptus (de Freitas et al. 2020), and reaction norms were constructed in arabica coffee (Bertrand et al. 2015) and used with random regression for GEI analysis in conventional eucalyptus breeding (Alves et al. 2020).

Beyond single-locus genotype data

Different types of molecular information can now be exploited by the GS model, which could lead to an increase in the accuracy of predictions by better modeling the genotype–phenotype relationship (Fig. 1).

The use of haploblocks made of two or more adjacent SNPs instead of single SNPs was investigated for genomic predictions, as it could increase GS accuracy by better capturing identity-by-descent between individuals, giving higher LD between QTLs and haploblock alleles, or capturing epistatic effects between

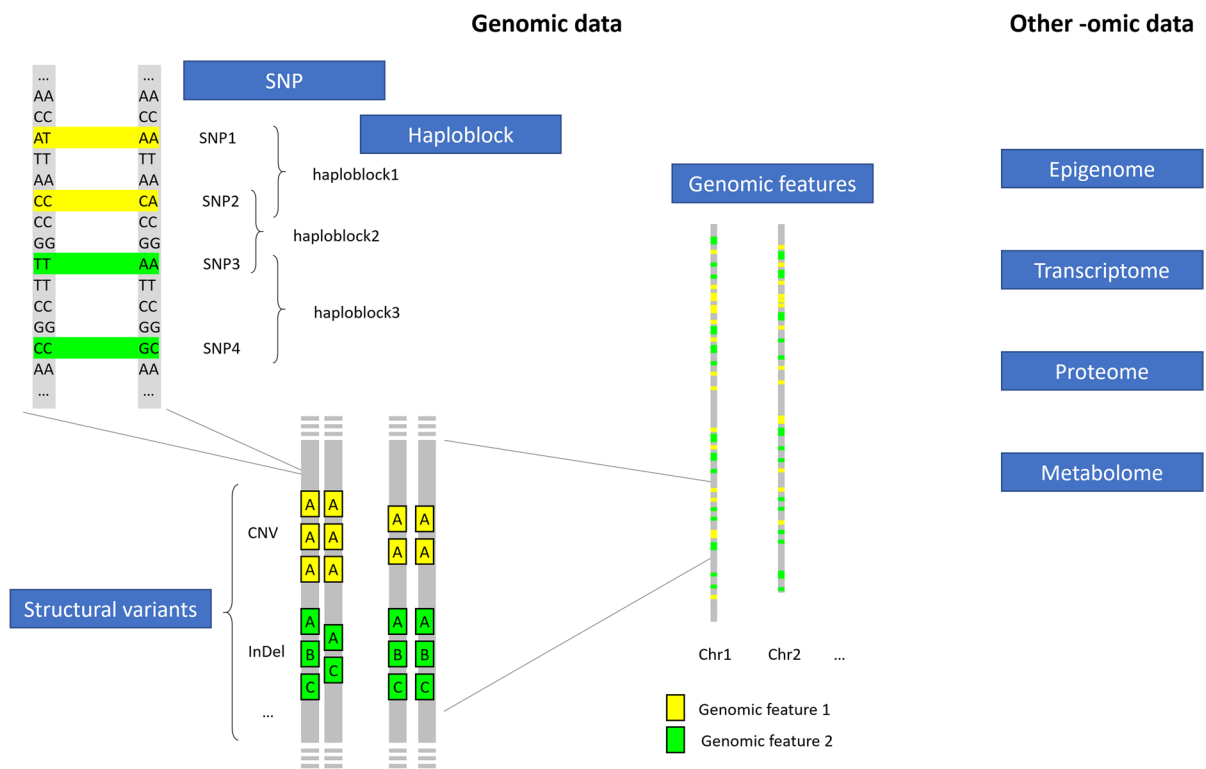


Fig. 1 Overview of possible molecular information for optimizing GS models. Genomic features can be defined in various ways: location in QTL, functional and structural annotations,

etc. (Sørensen et al. 2013). Two genomic features were considered here for illustration

SNPs in the same haploblock (Bhat et al. 2021; Goddard and Hayes 2007; Hess et al. 2017). Ballesta et al. (2019) explored the advantages of using haplotypic data for GS in *Eucalyptus globulus* and showed that prediction accuracy was significantly higher for low heritable traits when haploblocks were used instead of single SNPs. However, the relative efficiency of using haploblocks or single SNPs for genomic predictions is affected by many parameters, in particular the size of the training population, the level of LD, the method used to define the haploblocks, and the phasing accuracy (Bhat et al. 2021; Goddard and Hayes 2007; Hess et al. 2017). This aspect requires further investigation in tropical perennial crops and plantation trees.

The use of pangenomes is another possible avenue of GS research. Progress in sequencing techniques has enabled the comparison of individual genomes within species and shown that structural variations (SV) represent a significant proportion of polymorphism (Yuan et al. 2021). SVs consist of deletions, insertions, copy number variations, inversions, or translocations, with size > 50 bp. In particular, SVs include variations in gene presence/absence, with core genes that are found in all individuals and variable genes that are absent in some individuals. SVs cannot be represented by single reference genomes, and pangenomes are thus required to harness the whole genetic diversity of the breeding population (Bayer et al. 2021; Scossa et al. 2021). So far, very few studies have considered using structural variations for genomic predictions. In wheat, Würschum et al. (2017) obtained a slight increase in GS accuracy when markers specifically targeting a CNV contributing to the genetic control of the target trait were included in the model. Similarly, in maize and cattle, the use of CNV information in the GS model increased prediction accuracy in some cases (El Hamidi et al. 2018; Lyra et al. 2019). The use of SV information for genomic predictions deserves greater attention, and this will be greatly facilitated by pangenomes. Several reference genomes are already available for certain tropical perennial crops and plantation trees (e.g., cocoa and oil palm), and the next step should be the construction of pangenomes. The biggest impact could be on polyploid crops, such as bananas, as SV may represent an even higher proportion of polymorphisms in polyploids (Schissel et al. 2019).

Another way of improving GS accuracy is to incorporate existing information concerning polymorphisms, particularly that obtained from studies of QTL detection, in the prediction model (Xu et al. 2020). Different modeling approaches have been developed for this purpose, and their efficiency has been demonstrated in animal and plant studies, including temperate perennial fruit trees (Nsibi et al. 2020). However, very few studies have investigated this aspect in tropical perennial crops and plantation trees so far. In oil palm, Kwong et al. (2017a) applied RRBLUP using only SNPs with the highest GWAS association score, which made it possible to reduce marker density while achieving better or the same accuracy as using all the SNPs. A similar result was obtained in eucalyptus (Tan and Ingvarsson 2019). However, these approaches depend on a careful definition of the training and application populations. Thus, in cocoa, the inclusion of the SNPs detected by GWAS as fixed effects in the GS model did not improve prediction accuracies, which likely resulted from a too high genetic differentiation between the training and application populations, making the detected SNPs irrelevant (McElroy et al. 2018).

Incorporating endophenotypes, or intermediate phenotypes, in prediction models is another promising feature of GS research. Endophenotypes, and in particular transcriptomic and metabolomic data, have been used jointly with genomic data in a few crops (Scossa et al. 2021; Tong and Nikoloski 2021; Xu et al. 2020). These multi-omics prediction approaches are expected to better capture minor and non-additive effects and to better model the relationship between genotypes and phenotypes. Multi-omics prediction produced promising results in rice and maize, where they outperformed single-omic predictions. This requires specific statistical approaches, like machine learning (Montesinos-López et al. 2021; Tong and Nikoloski 2021). Investigating these aspects would be of interest to tropical perennial crops and plantation trees.

GS aided re-domestication and introgression breeding

Some perennial tropical crops have breeding populations with narrow genetic bases, and hence, only a fraction of the genetic diversity of the species is exploited, for instance, in *Coffea Arabica* (Tran

et al. 2016), cacao (Lanaud et al. 2001; Zhang and Motilal 2016), and rubber (Priyadarshan 2011). This usually resulted from choices and constraints dating back to the beginning of the breeding of these crops, or even before. In addition, the criteria originally used to select individuals might differ from the criteria that are of interest today, and current breeding populations may no longer correspond to current needs in terms of diversity. For example, in oil palm, the Deli breeding population, which today is used as one of the two heterotic populations mated to produce the vast majority of the oil palm cultivars, originated from four individuals collected in Africa and planted in Indonesia in 1848, decades before the establishment of the first commercial plantations (Corley and Tinker 2016). The other oil palm breeding populations derived from a small number of founders selected among individuals collected in restricted regions during prospecting, usually in the first half of the twentieth century. Although this led to reduced effective sizes (Cros et al. 2014), which is advantageous for GS accuracy, it constrains the long-term genetic gain. Also, for the La Mé oil palm breeding population, the founder individuals were selected in the 1920s, giving less importance to the proportion of pulp in the fruits than breeders do today (Cochard 2008). Although this has not prevented significant genetic progress (e.g., in oil palm, genetic progress is considered to be 1–1.5% per year (Rival and Levang 2014), and in rubber tree, yield increased from 500 kg ha⁻¹ in primary clones developed in the 1930–1960 period to 2500 kg ha⁻¹ in the best clones today (Priyadarshan 2011)), broader genetic diversity of the crops concerned would help maintain the rate of the genetic progress and likely increase it. This could be achieved through the re-domestication of existing crops (Tian et al. 2021), which consists in initiating breeding afresh from a renewed and broader diversity comprising ancestors and/or natural populations of existing crops. Introgression breeding could also play an important role in increasing genetic diversity by transferring exotic alleles from the related species of cultivated crops (Gramazio et al. 2021). GS is an attractive way of implementing these processes efficiently (Crossa et al. 2017). Indeed, re-domestication or introgression breeding of perennial tropical crops and plantation trees would normally require many decades of phenotypic selection, making GS a particularly attractive option. One example is already

available in a temperate perennial fruit tree, apple (Kumar et al. 2020), a study which suggested that, for the introgression of monogenic traits into a superior germplasm by backcrosses or pseudo-backcrosses, GS would be efficient for the background selection implemented among the individuals that inherited the trait of interest from the exotic donor germplasm, as it would accelerate the elimination of the unwanted alleles of the donor, compared to conventional phenotypic background selection. The use of GS for this purpose should be considered in perennial tropical crops and plantation trees where introgression breeding from wild species has already been shown to be of interest, including citrus, banana, and cacao (Scossa et al. 2016).

Combining profiles of predicted marker effects and targeted recombination

As mentioned above, one limiting factor in breeding perennial crops is the constrained size of the population of selection candidates, as the larger the population, the more exhaustive the search for elite individuals within the diversity generated by meiosis. GS makes it possible to increase the population of selection candidates by replacing phenotyping with genotyping. Controlling the gametes generated at meiosis could further increase the efficiency of the breeding scheme. This could be made possible by combining genome-wide profiles of marker effects estimated using GS models and targeted recombination (Bernardo 2017). The profiles of marker effects along the chromosomes of heterozygote individuals could be used to identify sites in the genome where recombinations would maximize the genetic value of their gametes by aggregating blocks of favorable alleles. Recombinations could be obtained at these sites through genome editing, and the progenies of the regenerated edited individuals were screened to identify the best ones. This approach has great potential to increase genetic progress (Bernardo 2017; Brandariz and Bernardo 2019b). Genome editing tools are under active development in perennial tropical crops and plantation trees, for example, in cacao (Fister et al. 2018) and oil palm (Yeap et al. 2021). However, further studies are required in these species to develop efficient, targeted recombination approaches and to evaluate the relative efficiency of breeding

schemes involving targeted recombinations and conventional schemes.

GS-based breeding consortia

Breeding for perennial crops is highly complex and very costly, and only limited resources are available for breeding many tropical perennials. Furthermore, as we have seen throughout this review, using GS requires expertise in a range of scientific and technical fields, including quantitative genetics, biostatistics, bioinformatics, genomics, computer programming, and, in particular, with the growing interest in machine learning, mathematics. GS also often requires a large training population which, in the context of climate change, will need to be evaluated in multiple environments. This puts tropical perennial crops in a completely different situation than many other crops including temperate cereals and legumes that can rely on a dynamic private sector to bring together the required human resources, phenotyping and genotyping capacities, etc. and to make rapid progress in innovative methods, resulting in the release of cultivars that have benefited from these methods. One possible solution for tropical perennial crops would be to strengthen international collaboration by sharing the efforts required for the practical implementation of GS, i.e., multi-environment phenotyping, high-throughput genotyping, and statistical analyses for genomic predictions. Sneller et al. (2021) called for the construction of GS-based breeding consortia, which would allow each member of a consortium to share the overall GS costs while predicting the genetic value of its selection candidates using a large training population comprising genetic material from all the consortium partners. Another advantage of such consortia would be the possibility to evaluate genetic material in different environments through the exchange of plant material among the consortium partners. Even so, there would have to be some relatedness between the plant material shared by the members of the consortium, and sufficient genotypes would have to be evaluated in different partners' environments (Sneller et al. 2021). Such a consortium is a possible solution for the implementation of GS for tropical perennial species on which, to our knowledge, no GS studies have been published so far, including coconut, papaya, avocado, mango, or teak, despite their major economic importance.

Projects in this sense are currently being set up for some perennial tropical crops and plantation trees, like coffee (World Coffee Research 2022), while others could emerge by building on existing networks, like MusaNet (<https://musanet.org/>) and CacaoNet (<https://www.cacaonet.org/>).

Conclusion

Genomic selection (GS) should revolutionize the breeding of perennial tropical crops and plantation trees as it has already produced promising results in terms of an increase in the rate of genetic progress. GS will (i) enable increased selection intensity and/or a shorter generation interval by replacing all or some phenotyping by genotyping in selected breeding cycles and (ii) increase accuracy for traits that are difficult to phenotype. Overall, the main factors that affect GS accuracy have been well studied in perennial tropical crops and plantation trees. However, the level of studies on GS varied in the following species: Some, like eucalyptus and oil palm, can be considered as models for GS including an in-depth assessment of its practical potential; in others, like banana and guava, GS studies were recently initiated, while in other species, like coconut, papaya, avocado, mango, and teak, despite their economic importance, no GS study has been conducted so far.

The results obtained in the plant and animal species where GS has been investigated to date suggest that optimal GS predictions could be achieved through joint analysis of all available information concerning genotype-to-phenotype relations, possibly including multiple omics and phenotypic data on multiple traits in several well-characterized environments, using prior information available on markers and all types of polymorphisms present in the populations concerned. For perennial crops, in which phenotyping is particularly complex and resource-consuming, there is an urgent need for increased international cooperation in the form of GS-based consortia to be able to gather such large datasets at a reasonable cost. The optimal implementation of GS will also require going beyond the standard GS technologies and methodologies used today. In particular, high-throughput phenotyping is a key approach to gathering the required amount of phenotypic data on such large populations at a reasonable rate and cost. Statistical methodologies able to handle

large multidimensional heterogeneous datasets are also required, and machine learning approaches are crucial, particularly artificial neural networks.

Future GS research in tropical perennial crops and plantation trees should systematically consider the use of single-step GBLUP when phenotypic data are available on ungenotyped individuals, the use of multivariate models when the traits of interest comprise correlated traits with contrasting levels of heritability, and random regression models for longitudinal traits. Training population optimization should also be undertaken. Targeted recombinations on sites identified based on the profiles of predicted marker effects should be investigated. Furthermore, GS has the potential to make re-domestication possible as well as to boost introgression breeding.

Acknowledgements The authors acknowledge the GENES program of the Intra-Africa Academic Mobility Scheme of the European Union for financial support (EU-GENES:2017-2552/001-001). The authors also thank Marie Denis, Gilles Trouche, André Clément-Demange, Angélique D'Hont, Dominique Dessauw, and Xavier Argout for discussions that improved the manuscript.

Author contribution EGS and DC carried out the literature review and wrote the manuscript, with help from WGA, NHB, NM, and JMB. All authors read and approved the final manuscript.

Funding This study was funded by the GENES Intra-Africa Academic Mobility Scheme of the European Union (EU-GENES:2017-2552/001-001) program, by CIRAD, and by a grant from PalmElit SAS.

Data availability Not applicable.

Declarations

Ethics approval Not applicable.

Consent for publication Not applicable.

Competing interests We declare that all authors do not have any kind of financial or non-financial interests that are directly or indirectly related to this review article.

References

- Akdemir D, Isidro-Sánchez J (2019) Design of training populations for selective phenotyping in genomic prediction. *Sci Rep* 9:1–15
- Alkimim ER, Caixeta ET, Sousa TV, Resende MDV, da Silva FL, Sakiyama NS, Zambolim L (2020) Selective efficiency of genome-wide selection in *Coffea canephora* breeding. *Tree Genet Genomes* 16:1–11
- Álvarez C, Pérez E, Cros E, Lares M, Assemat S, Boulanger R, Davrieux F (2012) The use of near infrared spectroscopy to determine the fat, caffeine, theobromine and (–)-epicatechin contents in unfermented and sun-dried beans of Criollo cocoa. *J near Infrared Spectrosc* 20:307–315
- Alves RS, de Resende MDV, Azevedo CF, de Rocha JRAS-CDO, Nunes ACP, Carneiro APS, dos Santos GA (2020) Optimization of Eucalyptus breeding through random regression models allowing for reaction norms in response to environmental gradients. *Tree Gen Genomes* 16:1–8
- Aneani F, Ofori-Frimpong K (2013) An analysis of yield gap and some factors of cocoa (*Theobroma cacao*) yields in Ghana. *Sustainable Agriculture Research* 2: 2:526–2016–37857
- Asaari MSM, Mertens S, Dhondt S, Inzé D, Wuyts N, Scheunders P (2019) Analysis of hyperspectral images for detection of drought stress and recovery in maize plants in a high-throughput phenotyping platform. *Comput Electron Agric* 162:749–758
- Ballesta P, Maldonado C, Pérez-Rodríguez P, Mora F (2019) SNP and haplotype-based genomic selection of quantitative traits in *Eucalyptus globulus*. *Plants (basel)* 8:331. <https://doi.org/10.3390/plants8090331>
- Bayer PE, Peteret J, Danilevicz MF, Anderson R, Batley J, Edwards D (2021) The application of pangenomics and machine learning in genomic selection in plants. *Plant Genome* 14:e20112
- Bernardo R (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci* 34:20–25
- Bernardo R (2020) Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity* 125:375–385. <https://doi.org/10.1038/s41437-020-0312-1>
- Bernardo R (2017) Prospective targeted recombination and genetic gains for quantitative traits in maize. *The Plant Genome* 10(2):1–9. <https://doi.org/10.3835/plantgenome2016.11.0118>
- Bertrand B, Bardil A, Baraille H, Dussert S, Doullbeau S, Dubois E, Severac D, Dereeper A, Etienne H (2015) The greater phenotypic homeostasis of the allopolyploid *Coffea arabica* improved the transcriptional homeostasis over that of both diploid parents. *Plant Cell Physiol* 56:2035–2051
- Bhat JA, Yu D, Bohra A, Ganie SA, Varshney RK (2021) Features and applications of haplotypes in crop breeding. *Commun Biol* 4:1–12
- Bhat JA, Ali S, Salgotra RK, Mir ZA, Dutta S, Jadon V, Tyagi A, Mushtaq M, Jain N, Singh PK, Singh GP, Prabhu KV (2016) Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front Genet* 7:221. <https://doi.org/10.3389/fgene.2016.00221>
- Blancón J, Dutartre D, Tixier M-H, Weiss M, Comar A, Praud S, Baret F (2019) A high-throughput model-assisted method for phenotyping maize green leaf area index

- dynamics using unmanned aerial vehicle imagery. *Front Plant Sci* 10:685
- Boote KJ, Jones JW, White JW, Asseng S, Lizaso JI (2013) Putting mechanisms into crop production models. *Plant, Cell Environ* 36:1658–1672
- Borges MVV, de Oliveira Garcia J, Batista TS, Silva ANM, Baio FHR, da Silva Junior CA, de Azevedo GB, de Oliveira Sousa Azevedo GT, Teodoro LPR, Teodoro PE (2021) High-throughput phenotyping of two plant-size traits of Eucalyptus species using neural networks. *J Forest Res* 33(2):591–599. <https://doi.org/10.1007/s11676-021-01360-6>
- Borrelli GM, Orrù L, Vita PD, Barabaschi D, Mastrangelo AM, Cattivelli L (2015) Chapter 18 - Integrated views in plant breeding: from the perspective of biotechnology. In: Sadras VO, Calderini DF (eds), *Crop Physiology*, 2nd edn, vol 2. Academic Press, San Diego pp 467–486. <https://doi.org/10.1016/B978-0-12-417104-6.00018-2>
- 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* (edib) 116:146–157. <https://doi.org/10.1038/hdy.2015.78>
- Brandariz SP, Bernardo R (2019a) Small ad hoc versus large general training populations for genomewide selection in maize biparental crosses. *Theor Appl Genet* 132:347–353. <https://doi.org/10.1007/s00122-018-3222-3>
- Brandariz SP, Bernardo R (2019b) Predicted genetic gains from targeted recombination in elite biparental maize populations. *Plant Genome* 12:180062. <https://doi.org/10.3835/plantgenome2018.08.0062>
- Brault C, Lazerges J, Doligez A, Thomas M, Ecartot M, Roumet P, Bertrand Y, Berger G, Pons T, François P, Le Cunff L, This P, Segura V (2022) Interest of phenomic prediction as an alternative to genomic prediction in grapevine. *Plant Methods* 18:108. <https://doi.org/10.1186/s13007-022-00940-9>
- Brauner PC, Müller D, Molenaar WS, Melchinger AE (2020) Genomic prediction with multiple biparental families. *Theor Appl Genet* 133:133–147
- Bustos-Korts D, Malosetti M, Chapman S, van Eeuwijk F (2016) Modelling of genotype by environment interaction and prediction of complex traits across multiple environments as a synthesis of crop growth modelling, genetics and statistics. *Crop Systems Biology* 55–82. https://doi.org/10.1007/978-3-319-20562-5_3
- Calleja-Rodríguez A, Pan J, Funda T, Chen Z, Baisson J, Isik F, Abrahamsson S, Wu HX (2020) Evaluation of the efficiency of genomic versus pedigree predictions for growth and wood quality traits in Scots pine. *BMC Genomics* 21:796. <https://doi.org/10.1186/s12864-020-07188-4>
- Cappa EP, de Lima BM, da Silva-Junior OB, Garcia CC, Mansfield SD, 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
- Cericola F, Lenk I, Fè D, Byrne S, Jensen CS, Pedersen MG, Asp T, Jensen J, Janss L (2018) Optimized use of low-depth genotyping-by-sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (*Lolium perenne* L.) *Front Plant Sci* 9:369. <https://doi.org/10.3389/fpls.2018.00369>
- Chattopadhyay K, Behera L, Bagchi TB, Sardar SS, Moharana N, Patra NR, Chakraborti M, Das A, Marndi BC, Sarkar A (2019) Detection of stable QTLs for grain protein content in rice (*Oryza sativa* L.) employing high throughput phenotyping and genotyping platforms. *Sci Rep* 9:1–16
- Chia Wong JA, Clement DPL, Mournet P, dos Santos Nascimento A, Solís Bonilla JL, Lopes UV, Pires JL, Gramacho KP (2022) A high-density genetic map from a cacao F2 progeny and QTL detection for resistance to witches' broom disease. *Tree Genet Genomes* 18:1–14
- Clark SA, Hickey JM, Daetwyler HD, van der Werf JH (2012) The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet Sel Evol* 44:1–9. <https://doi.org/10.1186/1297-9686-44-4>
- 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.).
- Collins AR (ed) (2007) Linkage disequilibrium and association mapping: analysis and applications, methods in molecular biology. Humana Press, p 376. <https://doi.org/10.1007/978-1-59745-389-9>
- Combs E, Bernardo R (2013) Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *The Plant Genome* 6(1):7. <https://doi.org/10.3835/plantgenome2012.11.0030>
- Corley RHV, Tinker PB (2016) *The oil palm*, 5th edn. Wiley-Blackwell, Chichester. UK
- Costa-Neto G, Fritsche-Neto R (2021) Enviromics: bridging different sources of data, building one framework. *Crop Breed Appl Biotechnol* 21:1–14. <https://doi.org/10.1590/1984-70332021v21Sa25>
- Costa-Neto G, Galli G, Carvalho HF, Crossa J, Fritsche-Neto R (2021) EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture. *G3* 11, jk040.
- 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* 127:981–994. <https://doi.org/10.1007/s00122-014-2273-3>
- 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 (2015) 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, 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* 18:1–17
- Cros D, Mbo-Nkoulou L, Bell JM, Oum J, Masson A, Soumahoro M, Tran DM, Achour Z, Le Guen V, Clement-Demange A (2019) Within-family genomic selection in rubber tree (*Hevea brasiliensis*) increases genetic gain for rubber production. *Ind Crops Prod* 138:111464
- Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de Campos los G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker

- S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK (2017) Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci* 22:961–975. <https://doi.org/10.1016/j.tplants.2017.08.011>
- Crossa J, Fritsche-Neto R, Montesinos-Lopez OA, Costa-Neto G, Dreisigacker S, Montesinos-Lopez A, Bentley AR (2021) The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. *Front Plant Sci* 12:651480. <https://doi.org/10.3389/fpls.2021.651480>
- Daetwyler HD, Calus MPL, Pong-Wong R, Campos G, de Hickey los JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193:347–365. <https://doi.org/10.1534/genetics.112.147983>
- Daval A, Pomiès V, Le Squin S, Denis M, Riou V, Breton F, Bink M, Cochard B, Jacob F, Billotte N (2021) In silico mapping in an oil palm breeding program reveals a quantitative and complex genetic resistance to *Ganoderma boninense*. *Mol Breed* 41(9):1–18. <https://doi.org/10.1007/s11032-021-01246-9>
- de Campos los G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193:327–345. <https://doi.org/10.1534/genetics.112.143313>
- de Freitas ECS, de Paiva HN, Neves JCL, Marcatti GE, Leite HG (2020) Modeling of eucalyptus productivity with artificial neural networks. *Ind Crops Prod* 146:112149
- de Moraes BFX, dos Santos RF, de Lima BM, Aguiar AM, Missiaggia AA, da Costa Dias D, Rezende GDPS, Gonçalves FMA, Acosta JJ, Kirst M (2018) Genomic selection prediction models comparing sequence capture and SNP array genotyping methods. *Mol Breeding* 38:1–14
- de Peixoto LA, Laviola BG, Alves AA, Rosado TB, Bhering LL (2017) Breeding *Jatropha curcas* by genomic selection: a pilot assessment of the accuracy of predictive models. *PLOS One* 12:e0173368. <https://doi.org/10.1371/journal.pone.0173368>
- de Souza LM, dos Santos LHB, Rosa JRBF, da Silva CC, Mantello CC, Conson ARO, Scaloppi EJJ, Fialho J de F, de Moraes MLT, Gonçalves P de S, Margarido GRA, Garcia AAF, Le Guen V, de Souza AP (2018) Linkage disequilibrium and population structure in wild and cultivated populations of rubber tree (*Hevea brasiliensis*). *Front Plant Sci* 9:815. <https://doi.org/10.3389/fpls.2018.00815>
- Denis M, Bouvet J-M (2013) Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. *Tree Genet Genomes* 9(1):37–51. <https://doi.org/10.1007/s11295-012-0528-1>
- Denis M, Cros D, Cochard B, Camus-Kulandaivelu L, Durand-Gasselín T, Bouvet JM (2012) Potential of genomic selection in perennial crops: preliminary results in the context of Eucalyptus and oil palm breeding : P-180 [WWW Document]. Programme and book of abstracts of the 4th International Conference of Quantitative Genetics: Understanding Variation in Complex Traits, Edinburgh, UK. 17–22. <http://agritrop.cirad.fr/568293/> (Accessed 6 Jun 2019)
- dos Santos Scholz MB, Kitzberger CSG, Pereira LFP, Davrieux F, Pot D, Charmetant P, Leroy T (2014) Application of near infrared spectroscopy for green coffee biochemical phenotyping. *J Near Infrared Spectrosc* 22(6):411–421. <https://opg.optica.org/jnirs/abstract.cfm?URI=jnirs-22-6-411>
- 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* 13:74. <https://doi.org/10.1007/s11295-017-1158-4>
- Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet* 126:1–11
- El Hamidi AH, Utsunomiya YT, Xu L, Zhou Y, Neves HH, Carneiro R, Bickhart DM, Ma L, Garcia JF, Liu GE (2018) Genomic predictions combining SNP markers and copy number variations in Nelore cattle. *BMC Genomics* 19:1–8
- Elli E, Sentelhas P, Freitas C, Carneiro R, Alcarde Alvares C (2019) Assessing the growth gaps of Eucalyptus plantations in Brazil – magnitudes, causes and possible mitigation strategies. *For Ecol Manag* 451:117464. <https://doi.org/10.1016/j.foreco.2019.117464>
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379. <https://doi.org/10.1371/journal.pone.0019379>
- Falconer D, Mackay T (1996) Introduction to quantitative genetics. Essex, UK: Longman Group
- Fanelli Carvalho H, Galli G, Ventrone Ferrão LF, Vieira Almeida Nonato J, Padilha L, Perez Maluf M, Ribeiro Resende de Jr MF, Guerreiro Filho O, Fritsche-Neto R (2020) The effect of bienniality on genomic prediction of yield in arabica coffee. *Euphytica* 216:1–16
- FAO, 2015. World programme for the census of agriculture 2020: volume 1-Programme, concepts and definitions.
- Ferrão LFV, Ferrão RG, Ferrão MAG, Fonseca A, Carbonetto P, Stephens M, Garcia AAF (2019) Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. *Heredity (edinb)* 122:261–275. <https://doi.org/10.1038/s41437-018-0105-y>
- Fister AS, Landherr L, Maximova SN, Guiltinan MJ (2018) Transient expression of CRISPR/Cas9 machinery targeting TcNPR3 enhances defense response in Theobroma cacao. *Frontiers Plant Sci* 9:268. <https://doi.org/10.3389/fpls.2018.0026>
- Flint-Garcia SA, Thornsberry JM, Buckler ES IV (2003) Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* 54:357–374
- Fugeray-Scarbel A, Bastien C, Dupont-Nivet M, Lemarié S (2021) R2D2 Consortium. Why and how to switch to genomic selection: lessons from plant and animal breeding experience. *Front Genet* 12:1185
- Gianola D, Van Kaam JB (2008) Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178:2289–2303
- Goddard ME, Hayes BJ (2007) Genomic selection. *J Anim Breed Genet* 124:323–330. <https://doi.org/10.1111/j.1439-0388.2007.00702.x>
- Gois IB, Borém A, Cristofani-Yaly M, Resende MDV, Azevedo C, Bastianel M, Novelli V, Machado M (2016)

- Genome wide selection in citrus breeding. *Genet Mol Res* 15(4). <https://doi.org/10.4238/gmr15048863>
- Gramazio P, Prohens J, Toppino L, Plazas M (2021) Introgression breeding in cultivated plants. *Frontiers Plant Sci* 12:764533. <https://doi.org/10.3389/fpls.2021.764533>
- Grattapaglia D, Silva-Junior OB, Resende RT, Cappa EP, Müller BSF, Tan B, Isik F, Ratcliffe B, El-Kassaby YA (2018) Quantitative genetics and genomics converge to accelerate forest tree breeding. *Front Plant Sci* 9:1693. <https://doi.org/10.3389/fpls.2018.01693>
- Grattapaglia D, Resende MD (2011) Genomic selection in forest tree breeding. *Tree Genet Genomes* 7:241–255
- 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*, vol 1. Managing, Sequencing and mining genetic resources. Springer Netherlands, Dordrecht, pp 651–682. https://doi.org/10.1007/978-94-007-7572-5_26
- Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol Biol* 57:461–485. <https://doi.org/10.1007/s11103-005-0257-z>
- Gupta PK, Kumar J, Mir RR, Kumar A (2010) Marker-assisted selection as a component of conventional plant breeding. *Plant Breed Rev* 33:145–217. <https://doi.org/10.1002/9780470535486.ch4>
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 92:433–443. <https://doi.org/10.3168/jds.2008-1646>
- Hazel LN, Lush JL (1942) The efficiency of three methods of selection. *J Hered* 33:393–399. <https://doi.org/10.1093/oxfordjournals.jhered.a105102>
- Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. *Crop Sci* 49(1):1–12. <https://doi.org/10.2135/cropsci2008.08.0512>
- Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31:423–447. <https://doi.org/10.2307/2529430>
- Heslot N, Jannink J-L, Sorrells M (2015) Perspectives for genomic selection applications and research in plants. *Crop Sci* 55:1–12. <https://doi.org/10.2135/cropsci2014.03.0249>
- Hess M, Druet T, Hess A, Garrick D (2017) Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genet Sel Evol* 49:54. <https://doi.org/10.1186/s12711-017-0329-y>
- Hickey LT, Hafeez AN, Robinson H, Jackson SA, Leal-Bertioli SCM, Tester M, Gao C, Godwin ID, Hayes BJ, Wulff BBH (2019) Breeding crops to feed 10 billion. *Nat Biotechnol* 37:744–754. <https://doi.org/10.1038/s41587-019-0152-9>
- Hu T, Chitnis N, Monos D, Dinh A (2021) Next-generation sequencing technologies: an overview. *Hum Immunol* 82:801–811
- Huth NI, Banabas M, Nelson PN, Webb M (2014) Development of an oil palm cropping systems model: lessons learned and future directions. *Environ Model Softw* 62:411–419
- Imai A, Kuniga T, Yoshioka T, Nonaka K, Mitani N, Fukamachi H, Hiehata N, Yamamoto M, Hayashi T (2019) Single-step genomic prediction of fruit-quality traits using phenotypic records of non-genotyped relatives in citrus. *PLoS One* 14:e0221880. <https://doi.org/10.1371/journal.pone.0221880>
- Isidro J, Jannink J-L, Akdemir D, Poland J, Heslot N, Sorrells ME (2015) Training set optimization under population structure in genomic selection. *Theor Appl Genet* 128:145–158. <https://doi.org/10.1007/s00122-014-2418-4>
- Isidro y Sánchez J, Akdemir D, (2021) Training set optimization for sparse phenotyping in genomic selection: a conceptual overview. *Front Plant Sci* 12:715910. <https://doi.org/10.3389/fpls.2021.715910>
- Isik F (2014) Genomic selection in forest tree breeding: the concept and an outlook to the future. *New Forest* 45:379–401. <https://doi.org/10.1007/s11056-014-9422-z>
- Ithnin M, Xu Y, Marjuni M, Serdari NM, Amiruddin MD, Low E-TL, Tan Y-C, Yap S-J, Ooi LCL, Nookiah R (2017) Multiple locus genome-wide association studies for important economic traits of oil palm. *Tree Genet Genomes* 13:1–14
- Jamnadas R, McMullin S, Iiyama M, Dawson IK, Powell B, Termote C, Ickowitz A, Kehlenbeck K, Vinceti B, van Vliet N, Keding G, Stadlmayr B, Van Damme P, Carsan S, Sunderland T, Njenga M, Gyau A, Cerutti P, Schure J, Kouame C, Obiri BD, Ofori D, Agarwal B, Neufeldt H, Degrande A, Serban A (2016) 2. Understanding the Roles of Forests and Tree-based Systems in Food Provision. In: Mansourian S, Vira B, Wildburger C (eds) *Forests and food: addressing hunger and nutrition across sustainable landscapes*, OBP Collection. Open Book Publishers, Cambridge, pp 29–72
- Jannink J-L, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics* 9:166–177. <https://doi.org/10.1093/bfpg/elq001>
- Juliana P, Montesinos-López OA, Crossa J, Mondal S, González Pérez L, Poland J, Huerta-Espino J, Crespo-Herrera L, Govindan V, Dreisigacker S (2019) Integrating genomic-enabled prediction and high-throughput phenotyping in breeding for climate-resilient bread wheat. *Theor Appl Genet* 132:177–194
- Kitony JK, Sunohara H, Tasaki M, Mori J-I, Shimazu A, Reyes VP, Yasui H, Yamagata Y, Yoshimura A, Yamasaki M (2021) Development of an Aus-derived nested association mapping (Aus-NAM) population in rice. *Plants* 10:1255
- Kopitke PM, Menzies NW, Wang P, McKenna BA, Lombi E (2019) Soil and the intensification of agriculture for global food security. *Environ Int* 132:105078. <https://doi.org/10.1016/j.envint.2019.105078>
- Kumar S, Hilario E, Deng CH, Molloy C (2020) Turbocharging introgression breeding of perennial fruit crops: a case study on apple. *Hortic Res* 7:1–7
- Kwong QB, Teh CK, Ong AL, Heng HY, Lee HL, Mohamed M, Low JZ-B, Apparow S, Chew FT, Mayes S, Kulaveerasingham H, Tammi M, Appleton DR (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, Chew FT, Tammi M, Mayes S, Kulaveerasingham H, Yeoh SH, Harikrishna JA, Appleton DR (2017) Genomic selection in commercial perennial crops: applicability and improvement in oil palm (*Elaeis*

- guineensis Jacq). *Scientific Reports* 7:2872. <https://doi.org/10.1038/s41598-017-02602-6>
- Kwong QB, Teh CK, Ong AL, Chew FT, Mayes S, Kulaveerasingam H, Tammi M, Yeoh SH, Appleton DR, Harikrishna JA (2017) Evaluation of methods and marker systems in genomic selection of oil palm (*Elaeis guineensis* Jacq). *BMC Genetics* 18:107. <https://doi.org/10.1186/s12863-017-0576-5>
- LaFramboise T (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res* 37:4181–4193. <https://doi.org/10.1093/nar/gkp552>
- Lanaud C, Motamayor JC, Risterucci A-M (2001) Implications of new insight into the genetic structure of *Theobroma cacao* L. for breeding strategies. In: Bekele F, End M, Eskes AB (eds) *Proceeding of the international workshop on new technologies and cacao breeding*, pp 89–107. Kota Kinabalu, Sabah 2001. <https://agritrop.cirad.fr/476853>
- Lebedev VG, Lebedeva TN, Chernodubov AI, Shestibratov KA (2020) Genomic selection for forest tree improvement: methods, achievements and perspectives. *Forests* 11:1190. <https://doi.org/10.3390/f11111190>
- Li Y, Suontama M, Burdon RD, Dungey HS (2017) Genotype by environment interactions in forest tree breeding: review of methodology and perspectives on research and application. *Tree Genet Genomes* 13:1–18
- Lin Z, Hayes B, Daetwyler H (2014) Genomic selection in crops, trees and forages: a review. *Crop Pasture Sci* 65:1177–1191
- Liu X, Wang H, Wang H, Guo Z, Xu X, Liu J, Wang S, Li W-X, Zou C, Prasanna BM, Olsen MS, Huang C, Xu Y (2018) Factors affecting genomic selection revealed by empirical evidence in maize. *Crop J* 6:341–352. <https://doi.org/10.1016/j.cj.2018.03.005>
- Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink JL (2011) Genomic selection in plant breeding. *Knowl Prospects ADVANCES IN AGRONOMY* 110:77–123. <https://doi.org/10.1016/B978-0-12-385531-2.00002-5>
- Lourenco D, Legarra A, Tsuruta S, Masuda Y, Aguilar I, Misztal I (2020) Single-step genomic evaluations from theory to practice: using SNP chips and sequence data in BLUPF90. *Genes* 11:790
- Lyra DH, Galli G, Alves FC, Granato ÍSC, Vidotti MS, e Sousa MB, Morosini JS, Crossa J, Fritsche-Neto R (2019) Modeling copy number variation in the genomic prediction of maize hybrids. *Theor Appl Genet* 132:273–288
- Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* 12:57–63
- Maldonado C, Mora F, Contreras-Soto R, Ahmar S, Chen J-T, do Amaral Júnior AT, Scapim CA (2020) Genome-wide prediction of complex traits in two outcrossing plant species through deep learning and Bayesian regularized neural network. *Frontiers Plant Sci* 11:1734
- Malosetti M, Bustos-Korts D, Boer MP, van Eeuwijk FA (2016) Predicting responses in multiple environments: issues in relation to genotype × environment interactions. *Crop Sci* 56:2210–2222
- Marchal A, Legarra A, Tisne S, Carasco-Lacombe C, Manez A, Suryana E, Omoré A, Nouy B, Durand-Gasselin T, Sánchez L (2016) Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq) progeny tests. *Molecular Breeding* 36:2
- Marchal A, Schlichting CD, Gobin R, Balandier P, Millier F, Muñoz F, Pâques LE, Sánchez L (2019) Deciphering hybrid larch reaction norms using random regression. *G3: Genes, Genomes, Genetics* 9:21–32
- McElroy MS, Navarro AJR, Mustiga G, Stack C, Gezan S, Peña G, Sarabia W, Saquicela D, Sotomayor I, Douglas GM, Migicovsky Z, Amores F, Tarqui O, Myles S, Motamayor JC (2018) Prediction of cacao (*Theobroma cacao*) resistance to *Moniliophthora* spp. diseases via genome-wide association analysis and genomic selection. *Front Plant Sci* 9:343. <https://doi.org/10.3389/fpls.2018.00343>
- Merrick LF, Herr AW, Sandhu KS, Lozada DN, Carter AH (2022) Optimizing plant breeding programs for genomic selection. *Agronomy* 12:714
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genet* 157:1819–1829
- Minamikawa MF, Nonaka K, Kaminuma E, Kajiya-Kanegae H, Onogi A, Goto S, Yoshioka T, Imai A, Hamada H, Hayashi T (2017) Genome-wide association study and genomic prediction in citrus: potential of genomics-assisted breeding for fruit quality traits. *Sci Rep* 7:1–13
- Momen M, Mehrgardi AA, Sheikhi A, Kranis A, Tusell L, Morota G, Rosa GJM, Gianola D (2018) Predictive ability of genome-assisted statistical models under various forms of gene action. *Sci Rep* 8:12309. <https://doi.org/10.1038/s41598-018-30089-2>
- Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JW, Fajardo-Flores SB, Gaytan-Lugo LS, Santana-Mancilla PC, Crossa J (2021) A review of deep learning applications for genomic selection. *BMC Genomics* 22:1–23
- Morota G, Gianola D (2014) Kernel-based whole-genome prediction of complex traits: a review. *Front Genet* 5:363
- Mphahlele MM, Isik F, Hodge GR, Myburg AA (2021) Genomic breeding for diameter growth and tolerance to leptocybe gall wasp and botryosphaeria/teratosphaeria fungal disease complex in *Eucalyptus grandis*. *Front Plant Sci* 12:228. <https://doi.org/10.3389/fpls.2021.638969>
- Mrode R, Ojango JMK, Okeyo AM, Mwacharo JM (2019) Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: current status and future prospects. *Front Genet* 9:694. <https://doi.org/10.3389/fgene.2018.00694>
- Mrode RA (2014) *Linear models for the prediction of animal breeding values*, 2nd edn. CABI International, Wallingford, Oxon, pp 235–245
- Müller BSF, Neves LG, de Almeida Filho JE, Resende MFR, Muñoz PR, dos Santos PET, Filho EP, Kirst M, Gratapaglia D (2017) Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of *Eucalyptus*. *BMC Genomics* 18:524. <https://doi.org/10.1186/s12864-017-3920-2>
- Munyengwa N, Le Guen V, Bille HN, Souza LM, Clément-Demange A, Mournet P, Masson A, Soumahoro M, Kouassi D, Cros D (2021) Optimizing imputation of

- marker data from genotyping-by-sequencing (GBS) for genomic selection in non-model species: rubber tree (*Hevea brasiliensis*) as a case study. *Genomics* 113:655–668. <https://doi.org/10.1016/j.ygeno.2021.01.012>
- Ni G, Cavero D, Fangmann A, Erbe M, Simianer H (2017) Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. *Genet Sel Evol* 49(1):1–14. <https://doi.org/10.1186/s12711-016-0277-y>
- Nielsen NH, Jahoor A, Jensen JD, Orabi J, Cericola F, Edriss V, Jensen J (2016) Genomic prediction of seed quality traits using advanced barley breeding lines. *PLoS One* 11(10). <https://doi.org/10.1371/journal.pone.0164494>
- Nsibi M, Gouble B, Bureau S, Flutre T, Sauvage C, Audergon J-M, Regnard J-L (2020) Adoption and optimization of genomic selection to sustain breeding for apricot fruit quality. *G3: Genes, Genomes Genet* 10:4513–4529
- Nyine M, Uwimana B, Blavet N, Hřibová E, Vanrespaille H, Batte M, Akech V, Brown A, Lorenzen J, Swennen R, Doležel J (2018) Genomic prediction in a multiploid crop: genotype by environment interaction and allele dosage effects on predictive ability in banana. *Plant Genome* 11:170090. <https://doi.org/10.3835/plantgenome2017.10.0090>
- Oliveira H, Brito L, Lourenco D, Silva F, Jamrozik J, Schaeffer L, Schenkel F (2019) Invited review: advances and applications of random regression models: from quantitative genetics to genomics. *J Dairy Sci* 102:7664–7683
- Paludeto JGZ, Grattapaglia D, Estopa RA, Tambarussi EV (2021) Genomic relationship-based genetic parameters and prospects of genomic selection for growth and wood quality traits in *Eucalyptus benthamii*. *Tree Genet Genomes* 17:1–20
- Peixoto MA, Alves RS, Coelho IF, Evangelista JSPC, de Resende MDV, de Rocha JRDOASC, e Silva FF, Laviola BG, Bhering LL (2020) Random regression for modeling yield genetic trajectories in *Jatropha curcas* breeding. *Plos one* 15:e0244021
- Persa R, de Oliveira Ribeiro PC, Jarquin D (2021) The use of high-throughput phenotyping in genomic selection context. *Crop Breed App Biotechnol* 21:1–11. <https://doi.org/10.1590/1984-70332021v21Sa19>
- Pirker J, Mosnier A, Kraxner F, Havlík P, Obersteiner M (2016) What are the limits to oil palm expansion? *Glob Environ Chang* 40:73–81. <https://doi.org/10.1016/j.gloenvcha.2016.06.007>
- Pootakham W, Jomchai N, Ruang-areerate P, Shearman JR, Sonthirod C, Sangsrakru D, Tragoonrung S, Tangphat-sornruang S (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. <https://doi.org/10.1016/j.ygeno.2015.02.002>
- Priyadarshan P (2011) *Biology of Hevea rubber*. Springer
- 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
- Rambolarimanana T, Ramamonjisoa L, Verhaegen D, Tsy J-MLP, Jacquin L, Cao-Hamadou T-V, Makouanzi G, Bouvet J-M (2018) Performance of multi-trait genomic selection for *Eucalyptus robusta* breeding program. *Tree Genet Genomes* 14:1–13
- Resende MDV, Resende MFR, Sansaloni CP, Petroli CD, Missiaggia AA, Aguiar AM, Abad JM, Takahashi EK, Rosado AM, Faria DA, Pappas GJ, Kilian A, Grattapaglia D (2012) Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol* 194:116–128. <https://doi.org/10.1111/j.1469-8137.2011.04038.x>
- Resende RT, Resende MDV, Silva FF, Azevedo CF, Takahashi EK, Silva-Junior OB, Grattapaglia D (2017) Assessing the expected response to genomic selection of individuals and families in *Eucalyptus* breeding with an additive-dominant model. *Heredity (edinb)* 119:245–255. <https://doi.org/10.1038/hdy.2017.37>
- Resende RT, Piepho H-P, Rosa GJ, Silva-Junior OB, de Resende MDV, Grattapaglia D (2021) Environments in breeding: applications and perspectives on envirotypic-assisted selection. *Theor Appl Genet* 134:95–112
- Reyes VP, Angeles-Shim RB, Mendioro MS, Manuel M, Carmina C, Lapis RS, Shim J, Sunohara H, Nishiuchi S, Kikuta M (2021) Marker-assisted introgression and stacking of major QTLs controlling grain number (Gn1a) and number of primary branching (WFP) to NERICA cultivars. *Plants* 10:844
- Rincent R, Kuhn E, Monod H, Oury F-X, Rousset M, Allard V, Le Gouis J (2017) Optimization of multi-environment trials for genomic selection based on crop models. *Theor Appl Genet* 130:1735–1752
- Rincent R, Charpentier J-P, Faivre-Rampant P, Paux E, Le Gouis J, Bastien C, Segura V (2018) Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar. *G3: Genes, Genomes Genet* 8:3961–3972
- Rival A, Levang P (2014) Palms of controversies: oil palm and development challenges. Bogor: Center for International Forestry Research. <https://doi.org/10.17528/cifor/004860>
- Robertsen CD, Hjortshøj RL, Janss LL (2019) Genomic Selection in Cereal Breeding. *Agronomy* 9:95. <https://doi.org/10.3390/agronomy9020095>
- Romero Navarro JA, Phillips-Mora W, Arciniegas-Leal A, Mata-Quirós A, Haiminen N, Mustiga G, Livingstone Iii D, van Bakel H, Kuhn DN, Parida L, Kasarskis A, Motamayor JC (2017) Application of genome wide association and genomic prediction for improvement of cacao productivity and resistance to black and frosty pod diseases. *Front Plant Sci* 8:1905. <https://doi.org/10.3389/fpls.2017.01905>
- Röös E, Bajželj B, Smith P, Patel M, Little D, Garnett T (2017) Greedy or needy? Land use and climate impacts of food in 2050 under different livestock futures. *Glob Environ Chang* 47:1–12. <https://doi.org/10.1016/j.gloenvcha.2017.09.001>
- Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94(3):441–448
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* 74:5463–5467. <https://doi.org/10.1073/pnas.74.12.5463>

- Schiessl S-V, Katche E, Ihien E, Chawla HS, Mason AS (2019) The role of genomic structural variation in the genetic improvement of polyploid crops. *The Crop Journal* 7:127–140
- Scossa F, Brotman Y, de e Lima FA, Willmitzer L, Nikoloski Z, Tohge T, Fernie AR (2016) Genomics-based strategies for the use of natural variation in the improvement of crop metabolism. *Plant Sci* 242:47–64
- Scossa F, Alseekh S, Fernie AR (2021) Integrating multi-omics data for crop improvement. *J Plant Physiol* 257:153352
- Silva FA, Viana AP, Corrêa CCG, Santos EA, Oliveira JAVS, Andrade JDG, Ribeiro RM, Glória LS (2021) Bayesian ridge regression shows the best fit for Ssr markers in *Psidium guajava* among Bayesian models. *Sci Rep* 11(1): 1–11. <https://doi.org/10.1038/s41598-021-93120-z>
- Silva-Junior OB, Faria DA, Grattapaglia D (2015) A flexible multi-species genome-wide 60K SNP chip developed from pooled resequencing of 240 *Eucalyptus* tree genomes across 12 species. *New Phytol* 206:1527–1540. <https://doi.org/10.1111/nph.13322>
- Slatkin M (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 9:477–485
- Sneller C, Ignacio C, Ward B, Rutkoski J, Mohammadi M (2021) Using Genomic selection to leverage resources among breeding programs: consortium-based breeding. *Agronomy* 11:1555
- Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE (2008) Genomic selection using different marker types and densities. *J Anim Sci* 86:2447–2454. <https://doi.org/10.2527/jas.2007-0010>
- Sørensen P, Edwards SM, Madsen P, Jensen P, Sørensen IF, de los Campos G, Sorensen D (2013) Genomic feature models: conference on genomics of common diseases. *Book of Abstracts. Conference on Genomics of Common Diseases*, Oxford, United Kingdom, 07/09/2013, p 68
- Sousa TV, Caixeta ET, Alkimim ER, Oliveira ACB, Pereira AA, Sakiyama NS, Zambolim L, Resende MDV (2019) Early selection enabled by the implementation of genomic selection in *Coffea arabica* breeding. *Front Plant Sci* 9:1934. <https://doi.org/10.3389/fpls.2018.01934>
- de Sousa IC, Nascimento M, Silva GN, Nascimento ACC, Cruz CD, de Almeida DP, Pestana KN, Azevedo CF, Zambolim L, Caixeta ET (2020) Genomic prediction of leaf rust resistance to *Arabica* coffee using machine learning algorithms. *Sci Agrár* 78(4):1–8. <https://doi.org/10.1590/1678-992X-2020-0021>
- Souza LM, Francisco FR, Gonçalves PS, Scaloppi Junior EJ, Le Guen V, Fritsche-Neto R, Souza AP (2019) Genomic selection in rubber tree breeding: a comparison of models and methods for managing G×E interactions. *Front Plant Sci* 10:1353. <https://doi.org/10.3389/fpls.2019.01353>
- Steiger JH (1980) Tests for comparing elements of a correlation matrix. *Psychol Bull* 87:245
- Sun D, Cen H, Weng H, Wan L, Abdalla A, El-Manawy AI, Zhu Y, Zhao N, Fu H, Tang J (2019) Using hyperspectral analysis as a potential high throughput phenotyping tool in GWAS for protein content of rice quality. *Plant Methods* 15:1–16
- Tan B, Grattapaglia D, Martins GS, Ferreira KZ, Sundberg B, Ingvarsson PK (2017) Evaluating the accuracy of genomic prediction of growth and wood traits in two *Eucalyptus* species and their F1 hybrids. *BMC Plant Biol* 17:110. <https://doi.org/10.1186/s12870-017-1059-6>
- Tan B, Grattapaglia D, Wu HX, Ingvarsson PK (2018) Genomic relationships reveal significant dominance effects for growth in hybrid *Eucalyptus*. *Plant Sci* 267:84–93. <https://doi.org/10.1016/j.plantsci.2017.11.011>
- Tan B, Ingvarsson PK (2019) Integrating genome-wide association mapping of additive and dominance genetic effects to improve genomic prediction accuracy in *Eucalyptus*. *bioRxiv* 15(2). <https://doi.org/10.1002/tpg2.20208>
- Thistlethwaite FR, El-Dien OG, Ratcliffe B, Klápště J, Porth I, Chen C, Stoehr MU, Ingvarsson PK, El-Kassaby YA (2020) Linkage disequilibrium vs pedigree: genomic selection prediction accuracy in conifer species. *PLOS ONE* 15:e0232201. <https://doi.org/10.1371/journal.pone.0232201>
- Tian Z, Wang J, Li J, Han B (2021) Designing future crops: challenges and strategies for sustainable agriculture. *Plant J* 105:1165–1178
- Tong H, Nikoloski Z (2021) Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data. *J Plant Physiol* 257:153354
- Tran HT, Lee LS, Furtado A, Smyth H, Henry RJ (2016) Advances in genomics for the improvement of quality in coffee. *J Sci Food Agric* 96:3300–3312
- Tyczewska A, Woźniak E, Gracz J, Kuczyński J, Twardowski T (2018) Towards food security: current state and future prospects of agrobiotechnology. *Trends Biotechnol* 36:1219–1229. <https://doi.org/10.1016/j.tibtech.2018.07.008>
- Uitdewilligen JG, Wolters A-MA, D’hoop BB, Borm TJ, Visser RG, Van Eck HJ (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS one* 8:e62355
- Van Eeuwijk FA, Bustos-Korts D, Millet EJ, Boer MP, Kruijer W, Thompson A, Malosetti M, Iwata H, Quiroz R, Kuppe C (2019) Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. *Plant Sci* 282:23–39
- VanRaden PM (2007) Genomic measures of relationship and inbreeding. *Interbull Bulletin* 37:33–33
- Varshney RK, Roorkiwal M, Sorrells ME (2017) Genomic selection for crop improvement: an introduction. In: Varshney RK, Roorkiwal M, Sorrells ME (eds) *Genomic selection for crop improvement: new molecular breeding strategies for crop improvement*. Springer International Publishing, pp 1–6. https://doi.org/10.1007/978-3-319-63170-7_1
- Voss-Fels KP, Cooper M, Hayes BJ (2019) Accelerating crop genetic gains with genomic selection. *Theor Appl Genet* 132:669–686. <https://doi.org/10.1007/s00122-018-3270-8>
- Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES (1998) Large-scale identification, mapping, and genotyping of

- single-nucleotide polymorphisms in the human genome. *Science* 280:1077–1082. <https://doi.org/10.1126/science.280.5366.1077>
- Wang N, Jassogne L, van Asten PJA, Mukasa D, Wanyama I, Kagezi G, Giller KE (2015) Evaluating coffee yield gaps and important biotic, abiotic, and management factors limiting coffee production in Uganda. *Eur J Agron* 63:1–11. <https://doi.org/10.1016/j.eja.2014.11.003>
- Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: current status and prospects. *Crop J* 6(4):330–340. <https://doi.org/10.1016/j.cj.2018.03.001>
- Wartha CA, Lorenz AJ (2021) Implementation of genomic selection in public-sector plant breeding programs: current status and opportunities. *Crop Breeding Appl Biotechnol* 21:1–19. <https://doi.org/10.1590/1984-70332021v21Sa28>
- Weir BS (1979) Inferences about linkage disequilibrium. *Biometrics* 35:235–254. <https://doi.org/10.2307/2529947>
- Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193:621–631. <https://doi.org/10.1534/genetics.112.146290>
- 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>
- Woittiez LS, van Wijk MT, Slingerland M, van Noordwijk M, Giller KE (2017) Yield gaps in oil palm: a quantitative review of contributing factors. *Eur J Agron* 83:57–77. <https://doi.org/10.1016/j.eja.2016.11.002>
- Wong CK, 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. <https://doi.org/10.1007/s00122-008-0715-5>
- World Coffee Research (2022) Innovea Global Coffee Breeding Network. <https://worldcoffeeresearch.org/programs/global-breeding-network>
- Wu D, Guo Z, Ye J, Feng H, Liu J, Chen G, Zheng J, Yan D, Yang X, Xiong X (2019) Combining high-throughput micro-CT-RGB phenotyping and genome-wide association study to dissect the genetic architecture of tiller growth in rice. *J Exp Bot* 70:545–561
- Würschum T, Longin CFH, Hahn V, Tucker MR, Leiser WL (2017) Copy number variations of CBF genes at the Fr-A2 locus are essential components of winter hardiness in wheat. *Plant J* 89:764–773
- Xu Y (2016) Envirotyping for deciphering environmental impacts on crop plants. *Theor Appl Genet* 129:653–673. <https://doi.org/10.1007/s00122-016-2691-5>
- Xu Y, Liu X, Fu J, Wang H, Wang J, Huang C, Prasanna BM, Olsen MS, Wang G, Zhang A (2020) Enhancing genetic gain through genomic selection: from livestock to plants. *Plant Communications* 1:100005. <https://doi.org/10.1016/j.xplc.2019.100005>
- Yeap W-C, Norkhairunnisa Che Mohd Khan, Norfadzilah Jamalludin, Muad MR, Appleton DR, Harikrishna Kulaveerasingam (2021) An efficient clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR-associated protein 9 mutagenesis system for oil palm (*Elaeis guineensis*). *Frontiers in Plant Science* 12
- Yuan Y, Bayer PE, Batley J, Edwards D (2021) Current status of structural variation studies in plants. *Plant Biotechnol J* 19:2153–2163
- Zhang D, Motilal L (2016) Origin, dispersal, and current global distribution of cacao genetic diversity. In: *Cacao Diseases*. Springer, pp 3–31. https://doi.org/10.1007/978-3-319-24789-2_1
- Zhou L, Holliday JA (2012) Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture. *BMC Genomics* 13:1–12
- Zuidema PA, Leffelaar PA, Gerritsma W, Mommer L, Anten NP (2005) A physiological production model for cocoa (*Theobroma cacao*): model presentation, validation and application. *Agric Syst* 84:195–225

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

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.