Mémoire présenté pour obtenir l'

HABILITATION A DIRIGER DES RECHERCHES

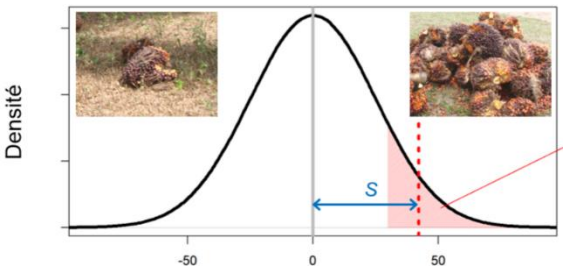
UNIVERSITE DE MONTPELLIER
Ecole Doctorale GAIA (N° 584)
(Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau)

Comment augmenter le rythme du progrès génétique pour les caractères complexes des plantes pérennes ? Exemple de l'intégration des prédictions génomiques dans les schémas d'amélioration du palmier à huile et de l'hévéa

par

David CROS

Chercheur Cirad
UMR 108 « Adaptation et Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales » (UMR AGAP Institut)



Soutenance le **16 décembre 2022** - Membres du Jury :
Mr Gilles CHARMET, rapporteur, Directeur de recherches, INRAE, UMR GDEC
Mme Laurence MOREAU, rapporteure, Directrice de recherches, INRAE, UMR GQE-Le Moulon
Mme Zulma VITEZICA, rapporteure, Professeur des universités, INPT, UMR GenPhySE
Mr Jacques DAVID, examinateur, Professeur des universités, Institut Agro Montpellier, UMR AGAP
Mme Christèle ROBERT-GRANIE, examinatrice, Directeur de recherches, INRAE, UMR GenPhySE

Table des matières

Table des figures

Remerciements

Je remercie les nombreux collègues, chercheurs, ingénieurs, techniciens, assistantes ou autres, avec qui j'ai collaboré depuis 2006, du Cirad et d'autres structures : de l'UPR 28 et/ou de PalmElit ; des équipes GFP et GS ; de mon équipe actuelle, GSP (en particulier des collectifs palmier à huile et hévéa) ; de l'UMR AGAP ; de l'UMR DIADE ; des plateaux de génotypage et de bioinformatique du Cirad ; de la direction régionale du Cirad à Yaoundé ; de l'INRAE (en particulier d'Orléans) ; du CRAPP de Pobè ; de l'université d'Abomey-Calavi ; de l'Université de Yaoundé 1 ; etc.

Je remercie enfin tout particulièrement Bruno Nouy, Albert Flori, Norbert Billotte, Léopoldo Sanchez, Jean-Marc Bouvet et David Lopez.

Je remercie aussi les étudiants, avec qui mes rapports furent aussi divers qu'enrichissants.

Déclaration d'intégrité scientifique

Je déclare avoir respecté, dans la conception et la rédaction de ce mémoire d'HDR, les valeurs et principes d'intégrité scientifique destinés à garantir le caractère honnête et scientifiquement rigoureux de tout travail de recherche, visés à l'article L.211-2 du Code de la recherche et énoncés par la Charte nationale de déontologie des métiers de la recherche et la Charte d'intégrité scientifique de l'Université de Montpellier. Je m'engage à les promouvoir dans le cadre de mes activités futures d'encadrement de recherche.

Liste des abréviations

AGC : aptitude générale à la combinaison
AGAP : Unité Mixte de Recherche « Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales » (UMR Institut AGAP)
ASC : aptitude spécifique à la combinaison
ANN : *artificial neural networks*
BIOS : département « Biologie des systèmes » (CIRAD)
BLUE : *best linear unbiased estimator*
BLUP : *best linear unbiased predictor*
CIRAD : centre de coopération internationale en recherche agronomique pour le développement
CGM : crop growth models
CRA-PP : centre de recherches agricoles sur les plantes pérennes
DL : déséquilibre de liaison
ENS : école Normale Supérieure (UY1)
FFB : *fresh fruit bunch* (poids total des régimes)
%FR : pourcentage de fruits dans le régime
GBS : *genotyping-by-sequencing*
GEBV : *genomic estimated breeding values*
GSP : équipe « Génome et sélection des pérennes » (AGAP)
GxE : interactions entre le génotype et l'environnement
%HP : pourcentage d'huile dans la pulpe fraîche des fruits
%HR : pourcentage d'huile dans la pulpe, ou taux d'extraction
HTP : *high-throughput phenotyping*
IFC : institut français du caoutchouc
LSCT : *large scale clonal trial*
NGS : *next generation sequencing*
NR : nombre de régimes
%PF : pourcentage de pulpe dans les fruits
PM : poids moyen des régimes
QTL : *quantitative trait locus*
NCM : *North Carolina model*
RRBLUP : *random-regression* ou *ridge-regression* BLUP
SALB : *South American leaf blight*
SET : *seedling evaluation trial*
SG : sélection génomique
SNP : *single nucleotide polymorphism*
SOCFINDO : société financière des caoutchoucs d'Indonésie
SOGB : société des caoutchoucs de Grand Bereby
SSCT : *small scale clonal trial*
SSR *: simple sequence repeats*
SVM : *support vector machine*
TPD : *tapping panel dryness*
UAC : université d'Abomey-Calavi (Bénin)
UY1 : université de Yaoundé 1 (Cameroun)

Avant-propos

J'ai été impliqué dans des recherches et j'ai eu des activités d'enseignement et d'encadrement d'étudiants pratiquement depuis mon recrutement en CDI au Cirad (Sept. 2006). Cependant, c'est principalement à partir de 2011 et du début de ma thèse de doctorat que j'ai conduit mes propres recherches. Elles ont essentiellement porté sur la sélection génomique, appliquée au palmier à l'huile puis aussi à l'hévéa. De façon ponctuelle, j'ai aussi réfléchi, voire travaillé, sur l'application de cette méthode chez d'autres espèces pérennes tropicales (bananier, eucalyptus, teck).

Dans ce mémoire, j'aborderai mes activités depuis 2011 sur la sélection génomique du palmier à huile et de l'hévéa, en présentant les aspects permettant de répondre aux objectifs de l'habilitation à diriger des recherches, à savoir « *la reconnaissance du haut niveau scientifique du candidat, du caractère original de sa démarche dans un domaine de la science, de son aptitude à maîtriser une stratégie de recherche dans un domaine scientifique ou technologique suffisamment large et de sa capacité à encadrer de jeunes chercheurs ».*

Partie 1. Curriculum vitae détaillé

1.1. Etat civil

David CROS

Adresse professionnelle :
CIRAD, UMR AGAP, TA A-108/1
34398 Montpellier Cedex 5
Tél. 04 67 61 44 50
Bât 1, bureau 14

1.2. Diplômes

- Montpellier Supagro **2014** : **Thèse de doctorat** « Etude des facteurs contrôlant l'efficacité de la sélection génomique chez le palmier à huile (*Elaeis guineensis* Jacq.) », mention très honorable, Directeurs de thèse : J. M. Bouvet (CIRAD), L. Sanchez (INRAE)
- Université Antilles Guyane **2005** : **DEA** Valorisation de la biodiversité tropicale
- ENSA Toulouse **2002** : **Ingénieur Agronome**, spécialisation amélioration des plantes

1.3. Compétences linguistiques

Français, Anglais, Espagnol : courant

1.4. Expérience professionnelle

01/2021 - présent : co-responsable de l'équipe « Génome et sélection des pérennes », CIRAD, UMR Institut AGAP / Montpellier
08/2020 - présent : **chercheur en amélioration génétique des plantes**, spécialiste du palmier à huile, CIRAD, UMR Institut AGAP / **Montpellier**

L'Unité Mixte de Recherche « Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales » (UMR Institut AGAP) rassemble le CIRAD, l'INRAE, l'Institut Agro et l'Université de Montpellier. Elle a été créée le 1er janvier 2011 et regroupe plus de 400 personnes. Elle fait partie du département BIOS du CIRAD. Elle vise à répondre aux grands enjeux de l'agriculture (sécurité alimentaire, durabilité des agro-systèmes et réduction de leur vulnérabilité, transition écologique). Elle mobilise une grande diversité de compétences, d'approches et de ressources dans les domaines de la biologie végétale, de la génétique, des biostatistiques, de l'amélioration des plantes, etc. Ses recherches portent sur une vingtaine d'espèces méditerranéennes et tropicales : espèces annuelles

autogames (riz, blé, sorgho, coton, arachide), espèces à reproduction contrainte (agrumes, racines et tubercules, bananiers, canne à sucre), et espèces pérennes (eucalyptus, palmier à huile, pommier, vigne, olivier, hévéa, cacao). Elle s'organise en quatre pôles scientifiques, (1) Diversités et dynamique évolutive des plantes cultivées, (2) Structure et dynamique des génomes, (3) Développement et fonctionnement des plantes et des peuplements », et (4) Déterminisme génétique et méthodologies de sélection ; constitués d'équipes, plateformes et plateaux (voir Annexe 1).

L'équipe « Génome et Sélection des Pérennes (GSP) » fait partie du pôle 4. Elle conduit des études génétiques et génomiques sur les caractères d'intérêt agronomique de plantes tropicales agro-industrielles (cacaoyer, caféier, hévéa, palmier à huile) et d'espèces forestières commerciales (eucalyptus, teck, acacia, etc.), notamment concernant le rendement, la qualité des produits et la résistance aux maladies. Son objectif est de contribuer à élucider les fondements génomiques de ces caractères et de proposer les bases d'une sélection et d'une création variétale plus performantes dans des systèmes de cultures plus respectueux de l'environnement. En 2022, GSP est composée de 34 agents permanents dont 30 chercheurs (7 HDR), et comprend la station Cirad de Combi (Guyane) et le plateau de mycologie d'AGAP.

En tant que chercheur de cette équipe, mes principales missions sont de **conduire des recherches sur la méthodologie de la sélection pour maximiser le rythme du progrès génétique**. Je m'intéresse tout particulièrement à la sélection génomique et à son application au palmier à huile et à l'hévéa. J'ai une activité d'encadrement d'étudiants, de montage de projets et, à la marge, d'enseignement.

Par ailleurs, en tant que co-responsable de GSP (avec David Lopez), j'ai des **activités managériales, administratives et de gestion (ressources humaines, budget).** Je m'investis aussi dans **l'orientation et l'animation des actions scientifiques de l'équipe**, en lien avec la direction d'AGAP. Je réalise ainsi la moitié des entretiens annuels des agents de l'équipe et, avec David Lopez, nous organisons des réunions d'équipe toutes les deux semaines, répondons aux demandes de la direction d'AGAP, relayons les informations au sein de l'équipe, etc. Nous définissons aussi nos besoins de compétences et les présentons à la direction de l'UMR. Nous avons obtenu fin 2021 le recrutement d'un chercheur sur la génétique de l'hévéa.

08/2015 - 07/2020 : chercheur en amélioration génétique des plantes, spécialiste du palmier à huile, CIRAD, UMR AGAP / **Université Yaoundé 1, Cameroun**

Dans le cadre de mon affectation au sein de l'Université de Yaoundé 1, mes activités étaient plus fortement tournées vers l'encadrement d'étudiants, la formation et l'enseignement, sur des aspects en lien avec la méthodologie de la sélection sur le palmier à huile et l'hévéa. Une partie de mes activités relevait de la construction de partenariats (implication dans des projets communs, etc.).

09/2011 - 07/2015 : chercheur en amélioration génétique des plantes, spécialiste du palmier à huile, CIRAD, **UMR AGAP** / **Montpellier**

Mes recherches sur l'optimisation de la méthodologie de la sélection du palmier à huile ont démarré en 2011, dans le cadre de ma thèse doctorat. Elles se sont accompagnées, dans une moindre mesure, d'une activité d'encadrement de stagiaires et d'enseignement.

09/2006 - 09/2011 : sélectionneur palmier à huile, CIRAD, UPR28 Amélioration génétique du palmier à huile / **CRAPP, Bénin**

J'ai eu la charge du suivi des activités de sélection, de production de semences et de recherche (essais en pépinière et au champ) sur l'amélioration du palmier à huile, menées conjointement par le Cirad / PalmElit et le CRAPP. J'ai initié des activités d'enseignement et d'encadrement de stagiaires.

10/2005 - 07/2006 : assistant sélectionneur concombres de serre type mini, **ENZA ZADEN**, **Espagne**

03/2003 - 09/2004 : volontaire civil à l'aide technique, programme ananas, **CIRAD-FLHOR**, **Martinique**

1.5. Stages dans des laboratoires français et/ou étrangers et collaborations productives

1.5.1. Stages

Janvier 2012, **INIA** (**Madrid**) : j'ai été accueilli pendant une semaine par J. Fernández pour démarrer l'adaptation de son logiciel MOLCOANC (reconstruction de pédigrée par *simulated annealing* à partir de données moléculaires) aux modes de reproduction des espèces végétales, et pour le rendre plus flexible. Un article a été publié (Cros et al., 2014).

1.5.2. Collaborations productives

J'ai collaboré à différents projets :
-projet **EU-GENES** (resp. E. Achigan-Dako, UAC/H. Ngalle, UY1), 2018-2022. Ce projet vise à augmenter le nombre de sélectionneurs en Afrique disposant d'une formation de haut niveau en génomique. Il propose des financements pour la réalisation de thèses de doctorats et de stages de masters pour des étudiants africains en mobilité entre pays d'Afrique. J'ai encadré 2 doctorants (E. Seyum et, en co-encadrement scientifique, L. Mbo-Nkoulou) et un étudiant master 2 (N. Munyengwa) dans ce projet. Plusieurs articles ont été publiés (Munyengwa et al., 2021 ; Seyum et al., 2022a ; Seyum et al., 2022b).

-**consortium "Oil Palm Genome Project"** (resp. N. Billotte, CIRAD). Ce consortium international rassemble le CIRAD, l'institut NEIKER (Espagne) et des acteurs de la recherche sur le palmier à huile (sociétés de plantation, centres de recherches publics). J'ai été impliqué dans les différents projets sur

la génétique et la génomique du palmier à huile portés par ce consortium : projet A (2010-2012), A+ (2014), B (2015-2019) et GeneExpress (2022-2025).

-**projet CETIC** (resp. C. Awono-Onana, M. Tchuente, UY1), 2015-2019. Ce projet, financé par la banque mondiale, a permis la mise en place d'un centre d'excellence basé à l'Université de Yaoundé 1 (Ecole Nationale Supérieure Polytechnique) et impliquant un réseau d'institutions en Afrique et hors du continent. Il comportait un volet sur la modélisation mathématique du vivant, avec des formations de master et thèse et des financements pour des projets de recherche. J'ai encadré un doctorant (A. Nyouma) et un stagiaire master 2 (E. Akpla), et co-porté un projet de recherche (ModStat) dans le cadre du CETIC. Plusieurs articles ont été publiés (Nyouma et al., 2019 ; Nyouma et al., 2022 ; Nyouma et al., 2020).

-**projet INRAE Breed2Last** (resp. L. Sanchez, INRAE), 2014-2019. J'ai été impliqué dans ce projet, intitulé « Optimal selection and mating accounting for non-additive genetic effects, genome diversity and Mendelian sampling terms » et financé par le métapogramme SELGEN de l'INRAE. J'ai encadré un stagiaire master 2 (B. Tchounke) et un article est en cours de finalisation (Tchounke et al., under review).

-**projets Institut Français du Caoutchouc** : IFC Création Variétale 3 (resp. A. Clément-Demange, CIRAD), 2019-2022, et IFC-CV4 (resp. V. Le Guen, CIRAD), 2022-2026. Ces projets portent sur la mise en œuvre d'un schéma d'amélioration génétique clonale de l'hévéa. J'ai encadré trois stagiaires master 2 dans ces projets (L. Mbo-Nkoulou, J. Oum II, N. Munyengwa), et j'encadrerai un doctorant (K. Daouda, 2022-2026). Deux articles sont parus (Cros et al., 2019 ; Munyengwa et al., 2021).

-**projets PalmElit** :
J'ai été impliqué dans différents projets conduits en partenariat avec PalmElit, filiale du CIRAD dédiée à l'amélioration génétique du palmier à huile, à la production et à la commercialisation de semences :
. FREEPALM (resp. D. Lopez, CIRAD), 2021-2022, qui vise à l'obtention d'une séquence du génome de haute qualité
. FRUITPALM_02 / THESE DOMONHEDO (resp. N Billotte, CIRAD), 2016-2019, sur la sélection pour la réduction de l'acidité de l'huile de palme. J'ai co-encadré un doctorant (H. Domonhedo), et deux articles ont été publiés (Domonhédo et al., 2018a ; Domonhédo et al., 2018b).
De façon moins récente, j'ai participé aux projets « PALMELIT Fusariose » (2009), « PALMELIT GxE » (2009) et « PALMELIT Lipalm » (2009).

J'ai apporté une contribution au montage du projet **BioTeak**, porté par Jean-Marc Gion (CIRAD) et qui vient d'être soumis à l'ANR. Je suis responsable d'un workpackage portant sur la sélection génomique appliquée au teck.

J'ai aussi collaboré au **réseau R2D2**, financé par le métaprogamme SELGEN de l'INRAE. Ce réseau est un groupe expert de chercheurs travaillant sur la sélection génomique chez différentes espèces animales et végétales. Il a notamment aboutit à la rédaction d'un article de synthèse bibliographique (R2D2 Consortium et al., 2021).

J'ai eu des collaborations ponctuelles avec des chercheurs étrangers :

(i) **Jesús Fernández (INIA, Madrid)**, dont j'ai adapté aux plantes le logiciel MOLCOANC (voir section 1.5.1)
(ii) **Pasi Rastas (Université d'Helsinki)**, qui a participé à une étude de génétique des populations qu'a réalisé un de mes doctorants, Essubalew Seyum (Seyum et al., 2022b).

J'ai mis en place en 2021 un **groupe de travail sur la simulation informatique des schémas d'amélioration**. Il implique 26 personnes (CIRAD et INRAE) et, pour l'instant, trois ateliers ont eu lieu pour recenser et comparer les outils existants, identifier les besoins puis commencer à tester les outils les plus prometteurs.

1.5.3. Missions à l'étranger

Entre 2012 et novembre 2022 j'ai effectué **21 missions à l'étranger** (Etats-Unis, Malaisie, Indonésie, Bénin, Cameroun, Côte d'ivoire) pour des conférences internationales, de l'appui scientifique, de l'enseignement et de l'encadrement d'étudiants (voir liste complète Annexe 2).

1.6. Obtention de contrats de recherche

J'ai obtenu et assumé la responsabilité des projets suivants :

Titre	Organisme	Année	Montant
CRESI SelGen3D	CIRAD	2021	25 KE
SELGEN_PALM	PalmElit	2021-2022	250 KE
ModStat*	CETIC (Banque Mondiale)	2017-2019	14 KE
METHODE DE SAM	PalmElit	2018-2020	408 KE
METHODE DE SAM	PalmElit	2017	140 KE
METHODE DE SAM	PalmElit	2016	139 KE
METHODE DE SAM	PalmElit	2015	148 KE
METHODE DE SAM	PalmElit	2014	101 KE
METHODE DE SAM	PalmElit	2013	103 KE
METHODE DE SAM	PalmElit	2012	102 KE
METHODE DE SAM	PalmElit	2011	87 KE
METHODE DE SAM	PalmElit	2010	102 KE
METHODE DE SAM	PalmElit	2009	78 KE
Total			1 697 KE

*co-responsable du projet avec Vivien Rossi (CIRAD)

1.7. Activités d'enseignement et formations

1.7.1. Enseignements

J'ai démarré des activités d'enseignement en 2008. J'ai dispensé l'**équivalent de plus de 70 jours d'enseignement de niveau master 1 et 2, en France, au Cameroun et au Bénin**.

2020 - 2022 (3 jours / an) **Breeding strategies for oil palm and rubber tree,** Master 1 emPlant, UniLasalle (Beauvais, France) – France
2018 - 2020 (1 jour / an) **Initiation à R,** Université Yaoundé 1, Master 2 Biologie des organismes végétaux - Cameroun

2014, 2016 (0.5 jour) **Amélioration génétique du palmier à huile,** Université Montpellier 2, Master 2 Biologie Fonctionnelle des Plantes/Biotechnologie des Plantes Tropicales - France

2016 - 2020 (3 jours / an), **Breeding plants for quantitative traits,** Université Yaoundé 1, master 2 Biologie des organismes végétaux – Cameroun

2016, 2017, 2019 (0.5 jour / an), **Application of mixed model analysis for crop improvement: prediction of genetic values with BLUP and Bayesian methods**, Université d'Abomey-Calavi, Master 2 Biostatistiques – Bénin

2016 - 2017 (0.5 jour / an) **Initiation à R,** Université Yaoundé 1, ENS – Cameroun

2008 - 2018 (4.5 jours / an) **Recherche en amélioration génétique, physiologie et agronomie du palmier à huile,** Université d'Abomey-Calavi / CRAPP (Pobè), Master Production Végétale - Bénin

1.7.2. Formations

J'ai dispensé plusieurs formations destinées à des doctorants et/ou chercheurs :

2016 (4 jours) **Analyses statistiques de données génomiques pour l'amélioration génétique : théorie et application sous R**, UY1 (Cameroun)
Pour cette formation, j'ai été porteur d'une réponse à l'action incitative du CIRAD « Formation collective au Sud », impliquant aussi S. Tisné et M. Denis, qui a permis d'obtenir le financement nécessaire à sa réalisation à Yaoundé.

2016 (3 jours) **Le logiciel R : comment débuter et s'en servir pour analyser des données de génétique et de génomique ?** IRAD (Cameroun) - Cameroun

J'ai par ailleurs dispensé plusieurs formations sur la sélection génomique dans le cadre des projets du consortium **OPGP** entre **2013** et **2019**, et j'ai participé en tant que formateur à l'**école-chercheurs** INRA « Sélection génomique - Théorie et mise en œuvre en relation avec les programmes d'amélioration » (**2013**, Bruz).

1.8. Activités d'encadrement scientifique

1.8.1. Encadrement de doctorants

1. KOUASSI Daouda, **2022-2025** « Optimisation de l'amélioration génétique de l'hévéa par l'utilisation des marqueurs moléculaires », Univ. Jean Lorougnon-Guede - Côte d'Ivoire, SoGB / projet IFC-CV4

2. TCHOUNKE Billy, **2018-2023** (48 mois) « Optimisation de la sélection génomique récurrente réciproque pour le rendement en huile de palme », UY1 – Cameroun, projets PalmElit « Méthodes de SAM / SelGen_Palm »,

3. SEYUM Essubalew, **2018-2022** (36 mois) « Genome properties of oil palm breeding populations and genomic predictions of hybrid performance », UY1 – Cameroun, projet EU-GENES

4. MBO-NKOULOU Luther, **2018-2022** (36 mois) « Assessment of banana accessions performances and genomic selection between drought and BSD contrasted conditions », UAC – Bénin, projet EU-GENES (co-encadrement scientifique, voir ci-dessous)

5. NYOUMA Achille, **2017-2021** (36 mois de thèse + 8 mois d CDD) « Extension des possibilités de la sélection génomique chez le palmier à huile par l'intégration de données moléculaires individuelles d'hybrides », Université Yaoundé 1 – Cameroun, projets PalmElit « Méthodes de SAM / SelGen_Palm », **soutenue le 3 novembre 2021**

Pour les thèses 2, 3 et 5, j'ai effectué l'encadrement scientifique des étudiants, l'encadrement administratif étant réalisé par un enseignant-chercheur de l'UY1 (Dr. Hermine Ngalle pour Essubalew Seyum, et Prof. Joseph M. Bell pour les autres). J'ai défini le sujet de ces thèses dans les projets PalmElit « Méthodes de SAM » et « SelGen_Palm » que je porte.

Pour la thèse 4, l'encadrement scientifique est partagé avec un enseignant-chercheur de l'UAC (Prof. Achigan-Dako), qui se charge aussi des aspects administratifs. Le sujet a été défini conjointement avec le Prof. Achigan-Dako.

Les différents financements obtenus (projets PalmElit que je porte et projet EU-GENES, porté par le Prof. Achigan-Dako de l'UAC) ont permis de prendre en charge l'ensemble des indemnités et des frais de fonctionnement des étudiants, dont des déplacements (Sénégal et France pour Achille Nyouma, France pour Billy Tchounke, Kenya pour Essubalew Seyum).

6. DOMONHEDO Hubert, 2015-2018 « Diversité génétique et déterminisme génétique de l'acidité dans les fruits mûrs chez le palmier à huile (Elaeis guineensis, Jacq.) », CRA-PP / UAC – Bénin, **soutenue le 1er février 2019**

J'étais encadrant d'une partie de cette thèse, dirigée par N. Billotte (CIRAD) et C. Ahanhanzo (UAC). Ce travail est éloigné de mes activités de recherche actuelles mais était dans le prolongement de mes activités lors de mon affectation au CRA-PP. J'ai essentiellement encadré le doctorant pour la réalisation d'un article de synthèse (Domonhédo et al., 2018a), la préparation du manuscrit de thèse,

les activités de terrain (choix du matériel végétal et phénotypage) et j'ai contribué à la rédaction d'un article (Domonhédo et al., 2018b).

1.8.2. Encadrement de Masters 2

J'ai encadré 14 étudiants de Master 2 (liste ci-dessous), lors de stages d'une durée de 6 mois, sauf pour Norman Munyengwa, sur 10 mois.

Nom	*Titre stage*	*Formation*	*Année*
Nathalie RENGASSAMY (co-encadrement avec X. ARGOUT)	Analyse des données de séquencage HiC et appel des *topologically associating domains* (TADs)	UM2, Master Sciences et Numérique pour la Santé	2021
Norman MUNYENGWA	Within-family genomic selection in rubber tree using genotyping-by-sequencing	University of Zimbabwe, Master of Plant Breeding	2019
Jean OUM II	Validation indépendante de la sélection génomique chez l'hévéa	UY1, Master Biologie des Organismes Végétaux	2017
Leopold POUOKAM	Prédiction par un modèle génomique de la valeur génétique pour la répartition de la production annuelle chez le palmier à huile	UY1, ENS, DIPES II* Mathématiques	2017
Evrard AKPLA	Comparaison de modèles univarié et bivarié pour prédire la production de croisements hybrides chez le palmier à huile	UAC, Master Biostatistiques	2017
Luther MBO-NKOULOU	Sélection génomique chez l'hévéa	UY1, Master Biologie des Organismes Végétaux	2016
Ferdinand DJOU MBOU	Indice de Gini et évaluation de la régularité de la reproduction chez le palmier à huile	UY1, ENS, DIPES II* Mathématiques	2016
Clément NGOMBO	Effet du type de marqueurs moléculaires sur l'estimation de paramètres génétiques : comparaison des SNP de génotypage par séquençage et des SSR	UY1, Master Biologie des Organismes Végétaux	2016
Billy TCHOUNKE	Comparaison de méthodes d'estimation de la valeur génétique basée sur le modèle linéaire mixte gaussien chez le palmier à huile	UY1, ENS, DIPES II* Mathématiques	2016
Alexandre MARCHAL	Sélection génomique multivariée chez le palmier à huile	UM2, Master Statistique Des Sciences De La Vie Et De La Sante	2014
Vincent SOUCHARD	Sélection du palmier à huile pour la régularité de la production de régimes	UM2, Master Statistique Des Sciences De La Vie Et De La Sante	2013
Magloire OTEYAMI	Sélection pour la résistance à la mineuse des feuilles dans le programme back cross interspécifique Elaeis oleifera x Elaeis guineensis	UAC, Master Production Végétale	2011

Hubert DOMONHEDO	Etude du déterminisme génétique de l'acidité des fruits chez le palmier à huile Elæis guineensis Jacq	UAC, Master Production Végétale	2010
Adolphe AGBO	Sélection pour la richesse en acide gras insaturés et la faible vitesse de croissance en hauteur dans le programme back cross Elaeis oleifera x Elaeis guineensis	UAC, Master Production Végétale	2009

*DIPES : diplôme de niveau bac+5 de l'Université de Yaoundé 1 (diplôme de professeur de l'enseignement secondaire)

1.8.3. Encadrement de chercheurs/ingénieurs en CDD

J'ai encadré en 2016 un étudiant camerounais, Billy Tchounke, en stage de DIPES II, une formation de niveau bac+5, mais ne permettant pas de s'inscrire en doctorat. Je l'ai ensuite recruté en CDD avec un contrat signé par la direction régionale du CIRAD à Yaoundé, pour une durée de 14 mois, pendant lesquels il a travaillé sur la gestion de la consanguinité dans le contexte de la sélection génomique chez le palmier à huile, avec un article (Tchounke et al., under review) et une présentation à la conférence PAG2020. Il a finalement effectué un Master 2 et je l'ai recruté en doctorat.

1.8.4. Publications impliquant les étudiants encadrés :

(**en gras souligné** = étudiants)

1. **Tchounke B.**, Sanchez L., Bell J.M. et Cros D., *Under review*. Mate selection: a useful approach to maximize genetic gain and control inbreeding in genomic and conventional oil palm (*Elaeis guineensis* Jacq.) hybrid breeding.

2. **Seyum E. G.**, Bille N. H., Abtew W. G., **Munyengwa N.**, Bell J. M., Cros D., Molecular breeding. Genomic selection in tropical perennial crops and plantation trees: a review. https://doi.org/10.1007/s11032-022-01326-4

3. **Seyum E. G.**, Bille N. H., Abtew W. G., ..., Cros D.; Journal of Applied Genetics. Genome properties of key oil palm (*Elaeis guineensis* Jacq.) breeding populations. https://doi.org/10.1007/s13353-022-00708-w

4. **Nyouma A.**, Bell J.M., Jacob F., …, Cros D., 2022. Improving the accuracy of genomic predictions in an outcrossing species with hybrid cultivars between heterozygote parents: case study of oil palm (*Elaeis guineensis* Jacq.). Mol. Genet. Genomics. https://doi.org/10.1007/s00438-022-01867-5

5. **Munyengwa N.**, Le Guen V., Bille H.N., Souza L.M., Clément-Demange A., …, 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.

6. **Nyouma A.**, Bell J.M., Jacob F., Riou V., Manez A., …, Cros D., 2020. Genomic predictions improve clonal selection in oil palm (*Elaeis guineensis* Jacq.) hybrids. Plant Science, 299: 110547.

7. Cros D., **Mbo-Nkoulou L.**, Bell J.M., **Oum J.**, Masson A. et al., 2019. Within-family genomic selection in rubber tree (*Hevea brasiliensis*) increases genetic gain for rubber production. Industrial Crops and Products, 138: 111464.

8. **Nyouma A.**, Bell J.M., Jacob F. et 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): 69.

9. Cros D., **Tchounke B.** et Nkague-Nkamba L., 2018. Training genomic selection models across several breeding cycles increases genetic gain in oil palm in silico study. Molecular Breeding, 38(7): 89.

10. **Domonhédo H.**, Cros D., Nodichao L., Billotte N. et Ahanhanzo C., 2018. Enjeux et amélioration de la réduction de l'acidité dans les fruits mûrs du palmier à huile, *Elaeis guineensis* Jacq. (synthèse bibliographique). Biotechnologie, Agronomie, Société et Environnement, 22(1)

11. **Marchal A.**, Legarra A., Tisné S., Carasco-Lacombe C., Manez A., …, Cros D., 2016. Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests. Molecular Breeding, 36(2): 1‑13.

1.9. Participation à des comités de thèse

1. **Geoffrey Haristoy**, UMR AGAP. « Prédiction et sélection génomique multi-caractère chez *Eucalyptus globulus* », **Oct 2019 – Sept. 2022**, Ecole Doctorale GAIA, Encadré par Jean-Marc GION (CIRAD) et Laurent BOUFFIER (INRAE)

2. **Alizarine Lorenzi**, UMR320 Génétique Quantitative Evolution Le Moulon, « Optimization of genomic selection for hybrids in a reciprocal selection program - Experimental evaluation and simulations on maize ». **Sept. 2020 - Août 2023.** Encadrée par Laurence Moreau (INRAE), Alain Charcosset (INRAE), Christina Lehermeier (RAGT2N).

1.10. Rapporteur de thèse

1. **Makobatjatji Mmoledi Mphahlele**, University of Pretoria. « Genomic breeding for accelerated improvement of growth, wood properties and plant defence in *Eucalyptus grandis* », **Mars 2022**.

1.11. Participation à des jurys de Master

1. **Héloïse Giraud, 2012**. Stage Master 2 « Haplotype reconstruction, application to the analysis of quantitative traits on maize », Ecole nationale supérieure agronomique de Montpellier, M2 Plant improvement and plant engineering / INRA Moulon

2. **Babacar Diouf, 2020**. Stage Ingénieur 3ème année « Précision de la sélection génomique dans des populations constituées de matériel élite et de ressources génétiques chez le pommier », Montpellier SupAgro Ingénieur systèmes agricoles et agroalimentaires durables au Sud / IRHS

1.12. Evaluation de projets

1. **OPTIMAGICS**, Optimal Mating with Genomic Selection, 2016, Soumis dans le cadre du Métaprogramme SelGen (INRAE)

1.13. Activités de referee

1. 2022, **Molecular Breeding** (sélection génomique sur des populations inter-spécifiques de palmier à huile)

2. 2020, **Industrial Crops and Products** (article de synthèse sur les approches moléculaires pour l'amélioration du palmier à huile)

3. 2019, **Industrial Crops and Products** (article sur la diversité génétique et la sélection dans une population de palmier à huile du Sénégal)

4. 2018, Revue Scientifique et Technique «**Forêt et Environnement du Bassin du Congo**» (comparaison d'hybrides d'eucalyptus plantés avec et sans contraintes nutritionnelles)

5. 2017, **PlosOne** (article sur la sélection génomique appliquée au palmier à huile)

1.14. Interventions dans des colloques et congrès

Principales interventions (présentations orales) :

1. Cros D., Tchounke B., Sanchez L**. 2020**. Inbreeding management and optimization of genetic gain with phenotypic and genomic selection in oil palm (*Elaeis guineensis*)[W776]. *Abstracts workshops of the PAG XXVIII*. **San Diego** : PAG, 1 p. Plant and Animal Genome. 28, 2020-01-11/2020-01-15, San Diego (Etats-Unis). https://plan.core-apps.com/pag_2020/abstract/821492dd-aec3-4064-a6c2-4b1db29b29ab

2. Cros D., Jacob F., Nyouma A., Tchounke B., Afandi D., Syahputra I., Cochard B**. 2019**. Advances in oil palm genomic selection. Kuala Lumpur : MPOB, 8 p. MPOB International Palm Oil Congress and Exhibition (PIPOC 2019), 2019-11-19/2019-11-21, **Kuala Lumpur** (Malaisie).

3. Cros D., Denis M., Sanchez L., Cochard B., Flori A., Durand-Gasselin T., Nouy B., Omoré A., Pomiès V., Riou V., Suryana E., Bouvet J.M. **2016**. Précision de prédiction génomique chez une plante pérenne : cas du palmier à huile. In : Deretz Séverine (ed.). *Sélection génomique : théorie et mise en oeuvre en relation avec les programmes d'amélioration*. Paris : INRA, p. 197-214 Ecoles-chercheurs INRA sur la sélection génomique 2013 - Théorie et mise en oeuvre en relation avec les programmes d'amélioration, 2013-09-23/2013-09-27, **Bruz**.

4. Bastien C., Cros D., This P. **2016**. Quelle place pour la sélection génomique chez les plantes pérennes ?. In : Deretz Séverine (ed.). *Sélection génomique : théorie et mise en oeuvre en relation avec les programmes d'amélioration*. Paris : INRA, p. 99-125 Ecoles-chercheurs INRA sur la sélection génomique 2013 - Théorie et mise en oeuvre en relation avec les programmes d'amélioration, 2013-09-23/2013-09-27, **Bruz**.

5. Cros D., Riou V., Tisne S., Sidibé-Bocs S., Ortega Abboud E., Argout X., Pomiès V., Nodichao L., Lubis Z., Cochard B., Durand-Gasselin T. **2016**. Empirical prediction accuracy of genomic selection between experimental designs and generations in oil palm. [W667]. San Diego : PAG, 1 p. Plant and Animal Genome Conference. 24, 2016-01-09/2016-01-13, **San Diego** (Etats-Unis). https://pag.confex.com/pag/xxiv/webprogram/Paper20286.html

6. Cros D., Denis M., Bouvet J.M., Sanchez L. **2015**. Genomic selection for heterosis without dominance in multiplicative traits: case study of bunch production in oil palm [W554]. *Plant and Animal Genomes Conference XXIII Conference, San diego, United States, San Diego, United States, January 10-14, 2015*. s.l. : s.n. Plant and Animal Genome Conference. 23, 2015-01-10/2015-01-14, **San Diego** (Etats-Unis). https://pag.confex.com/pag/xxiii/webprogram/Paper14693.html

7. Cros D., Denis M., Sanchez L., Cochard B., Flori A., Durand-Gasselin T., Nouy B., Omoré A., Pomies V., Riou V., Suryana E., Bouvet J.M. **2014**. Practical aspects of genomic selection in oil palm (*Elaeis guineensis*). In : ISOPB. *International Colloquium on Harnessing the Oil Palm Genome for Breeding, Bali, Indonesia, 16 june 2014*. s.l. : s.n., 2 p. International Colloquium on Harnessing the Oil Palm Genome for Breeding, 2014-06-16, **Bali** (Indonésie).

8. Cros D., Denis M., Sanchez L., Cochard B., Flori A., Durand-Gasselin T., Nouy B., Omoré A., Pomies V., Riou V., Suryana E., Bouvet J.M. **2014**. Genomic selection in oil palm (*Elaeis guineensis* Jacq.). In : IOPRI. *4th International Oil Palm Conference, Bali, Indonesia, 17-19 June 2014*. s.l. : s.n., 15 p. International Oil Palm Conference. 4, 2014-06-17/2014-06-19, **Bali** (Indonésie).

9. Cros D. **2013**. Factors controlling accuracy of genomic selection in oil palm (*Elaeis guineensis*) : W528. *Plant and Animal Genome XXI conference, San Diego, USA, January 12-16 2013,*. s.l. : s.n., 1 p. Plant and Animal Genome conference. 21, 2013-01-12/2013-01-16, **San Diego** (Etats-Unis). https://pag.confex.com/pag/xxi/webprogram/Paper8267.html

10. Cros D., Denis M., Sanchez L., Cochard B., Durand-Gasselin T., Bouvet J.M. **2012**. Genomic selection in small populations with reduced effective size: example of oil palm. *Tree Breeding, Genomics and Evolutionary Biology: New synergies to tackle the impact of climate change in the 21st century : Final Conference and Workshops of Noveltree Project, Helsinki and Vantaa, Finland, 6th-18th October 2012*. s.l. : s.n., 1 p. Final Conference and Workshops of Noveltree Project, 2012-10-16/2012-10-18, **Helsinki** (Finlande).

Partie 2. Synthèse des travaux de recherche

Dans cette seconde partie, je présenterai tout d'abord rapidement mes sujets d'étude et les données dont j'ai disposé (section 2.1). Je rappellerai ensuite le contexte dans lequel se sont inscrites mes recherches (2.2), en présentant le concept de sélection génomique (SG) et les deux plantes sur lesquelles j'ai conduit l'essentiel de mes travaux, le palmier à huile et l'hévéa. J'insisterai plus sur le palmier à huile, sur lequel je me suis investi davantage. Je détaillerai ensuite mes travaux sur la SG (2.3, 2.4 et 2.5), en les replaçant dans le contexte plus large des études traitant des mêmes sujets, réalisées par d'autres groupes sur les mêmes espèces ou sur d'autres espèces.

2.1. Introduction

2.1.1. Sujets d'étude

Depuis 2011, je conduis des recherches sur la SG, en travaillant essentiellement sur le cas du palmier à huile et de l'hévéa. J'ai eu **deux principaux thèmes de recherches** :
- la compréhension des **facteurs influençant la précision des prédictions génomiques**,
- la définition de **schémas d'amélioration intégrant les prédictions génomiques de façon à augmenter le rythme du progrès génétique**.

2.1.2. Approches mises en œuvre, outils utilisés et jeux de données

J'ai traité ces sujets avec deux grands types d'approches. Je me suis appuyé sur des **études empiriques**, conduites grâce à des données expérimentales collectées par nos partenaires (PalmElit, SOCFINDO, CRAPP pour le palmier à huile ; IFC pour l'hévéa). Cependant, la mise en œuvre pratique de l'amélioration génétique des plantes pérennes est complexe, et nécessite des investissements importants sur le long terme. Il est donc particulièrement difficile d'obtenir des données expérimentales pour évaluer le progrès génétique qu'offrirait la SG. Dans ce contexte, les **simulations informatiques** sont particulièrement adaptées (Hoban et al., 2012 ; Yuan et al., 2012). En complément des études empiriques, j'ai donc aussi conduit des études par simulation.

Les données expérimentales utilisées consistaient :

- pour le **palmier à huile**, en une **population complexe constituée d'environ 200 individus de chacun des groupes A et B**, avec des niveaux d'apparentements variés au sein des groupes (parents-enfants, plein-frères, demi-frères, cousins, …) et une structuration en termes de populations (Deli et Angola pour le groupe A, et différentes populations africaines pour le groupe B). Ces individus ont été testés sur descendance hybride, avec **plus de 500 croisements** répartis sur **deux sites expérimentaux** d'**Indonésie** (SOCFINDO) et comprenant environ 36 000 individus au total (Nyouma et al., 2022). Des observations étaient disponibles pour les composantes du rendement et la croissance en hauteur. Seule une partie de ces données étaient disponibles au début de mon doctorat, et les différentes études que j'ai publié sur ces sites expérimentaux ont donc été faites avec des jeux de données de taille croissante (Cros et al., 2017 ; Cros et al., 2015b ; Marchal et al., 2016 ; Nyouma et al., 2022 ; Nyouma et al., 2020). De la même façon, les données moléculaires disponibles ont évolué, avec initialement des **SSR** (Cros et al., 2015b ; Marchal et al., 2016), puis du **GBS** (Cros et al., 2017 ; Nyouma et al., 2022 ;

Nyouma et al., 2020) et finalement une **puce à SNP**. Les génotypages ont porté sur les individus A et B ainsi que sur 399 de leurs enfants hybrides. Pour l'étude de Nyouma et al. (2020), traitant de la sélection d'ortets, j'ai aussi utilisé les données moléculaires (GBS) et phénotypiques d'un essai clonal planté à la SOCFINDO. J'ai accédé aux données phénotypiques et acquis les données moléculaires dans le cadre de mes projets avec PalmElit « Méthodes de SAM » puis « 'SelGen_Palm ».

- pour l'**hévéa**, j'ai surtout travaillé sur une famille de **plein-frères**, comportant environ **300 clones** évalués dans **deux essais** clonaux à petite échelle (voir 2.2.5.g), installés sur deux sites de **Côte d'Ivoire** (SOGB). Les données phénotypiques étaient disponibles pour la production de latex et la teneur en saccharose. Les données de génotypage consistaient en des données **SSR**, puis **GBS**. Ces données ont été acquises dans le cadre du projet « IFC-CV3 », auquel je participe. J'ai aussi utilisé des données fournies par une chercheur de l'Université de Campinas (**Brésil**), portant sur une autre famille de plein-frères, avec environ **250 clones** génotypés par GBS et évalués pour la circonférence du tronc en essai clonal à petite échelle.

Pour les analyses, j'ai surtout utilisé le **logiciel R**, et en particulier les packages **ASReml-R** (Butler et al., 2009), **BGLR** (Pérez et de los Campos, 2013) et **rrBLUP** (Endelman, 2011). ASReml-R m'a servi pour les analyses de modèles linéaires mixtes avec la méthodologie BLUP (Annexe 3), avec des données de pédigrée ou de marqueurs (GBLUP, voir 2.2.3.c), en univarié ou multivarié (2.3.1.d). Pour les prédictions génomiques avec d'autres approches que le GBLUP, j'ai utilisé BGLR, qui propose les différentes méthodes de SG dites de l' « alphabet Bayésien » (BayesA, BayesB, BayesCπ, etc.) et le *reproducing kernel Hilbert spaces* (2.2.3.c). Pour appliquer la méthode de prédiction génomique RRBLUP chez l'hévéa, j'ai préféré le package rrBLUP car il est beaucoup plus rapide que l'équivalent Bayésien implémenté dans BGLR (*Bayesian random regression*), et gratuit, contrairement à ASReml-R. J'ai aussi eu ponctuellement recours à des logiciels dédiés, comme **Tassel 5 GBS** (Glaubitz et al., 2014) pour l'appel des SNP à partir de données brutes de séquences (voir 2.2.3.d), **Lep-Map3** (Rastas, 2017) pour la construction d'une carte génétique, et **Beagle** (Browning et al., 2018) et **LinkImputeR** (Money et al., 2017) pour l'imputation des génotypes manquants.

Pour les **simulations**, j'ai opté pour le palmier à huile pour une approche ***forward-in-time***. Elle est centrée sur les individus, à travers la simulation des haplotypes, de la méiose, de la sélection et des croisements (Figure 1). Elle peut être relativement lente et elle nécessite la connaissance des caractéristiques génétiques des populations de base à simuler, mais elle est capable de modéliser des scénarios complexes. Les simulations *forward-in-time* sont donc particulièrement adaptées pour étudier l'effet d'un nombre limité de générations de sélection artificielle chez des espèces bien caractérisées pour lesquelles les

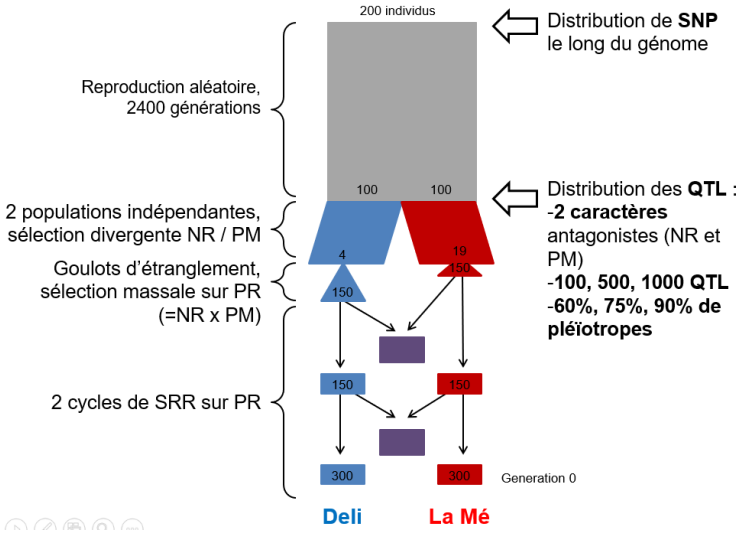


Figure 1 Procédure pour la simulation des populations de base de l'article Cros et al. (2017)

données disponibles permettent de définir les conditions génétiques initiales (Hoban et al., 2012 ; Yuan et al., 2012). Pendant mon doctorat, j'ai implémenté cette méthode dans R avec les fonctions fournies par le package HaploSim (Coster et Bastiaansen, 2010). Ce script a ensuite été complété pour d'autres études, en particulier dans le cadre du travail de Billy Tchounke. Pour l'hévéa, la simulation réalisée pour l'instant a porté sur une seule génération et a été faite plus simplement, en simulant des valeurs génétiques et phénotypiques, sans passer par la simulation d'haplotypes. Pour les deux espèces, les simulations ont été calibrées sur la base de paramètres génétiques calculés à partir des données empiriques mentionnées ci-dessus.

L'essentiel de mes analyses et des analyses de mes étudiants ont été effectuées sur un **serveur de calcul distant**, tout d'abord « sepang » puis « CC2 » au Cirad, et maintenant « meso@LR » de l'université de Montpellier (Figure 2). Le travail est réalisé en lignes de commandes ou via *RStudio server*.

Figure 2 Le cluster de calcul haute performance de l'Université de Montpellier (https://www.lemondeinformatique.fr/)

2.1.3. Organisation des recherches

Mes recherches depuis 2011 ont été essentiellement structurées par les **projets Cirad-PalmElit** que je porte sur la sélection génomique, puis les **projets IFC-CV** portés par mes collègues du collectif hévéa (André Clément-Demange, Vincent Le Guen).

Mon organisation a été dépendante de mon affectation, avec des conditions pendant les cinq années passées à Yaoundé très différentes de celles rencontrées à Montpellier.

Ainsi, une des raisons ayant motivé mon affectation à l'**université de Yaoundé 1** était qu'à ce moment-là il n'y avait pas sur place les compétences dont je disposais en génétique quantitative et sélection génomique, que je devais donc transférer. Il y avait par ailleurs un grand nombre d'étudiants, et proportionnellement peu d'offres de stages et de thèse en génétique. J'ai donc eu la possibilité d'encadrer de nombreux stagiaires de M2 et plusieurs doctorants. J'ai dû pour cela m'adapter et travailler avec les enseignants-chercheurs présents sur place (notamment les Prof. Bell et Ngalle), qui se sont chargés des aspects administratifs. Ils m'ont aussi aidé sur la sélection des étudiants, par exemple pour identifier les enseignements pertinents dans les relevés de note. Des difficultés techniques se sont posées, avec notamment un accès internet quasi-inexistant dans les locaux de l'université, alors que bon nombre de mes étudiants devaient travailler à distance sur le serveur de

calcul du Cirad. J'ai utilisé les financements de projets pour prendre en charge des modems et des achats mensuels de *data*. J'ai fait des économies d'échelles de mon temps en organisant des formations collectives pour les étudiants que j'encadrais simultanément, ou en les incluant dans les formations que je dispensais en dehors de mes enseignements de M2. J'ai aussi été confronté au fait que le niveau des étudiants était plus hétérogène que dans les formations que je connaissais avant, en particulier en termes de compétences en statistiques et en informatique. Je me suis alors appuyé sur mes enseignements pour identifier les étudiants avec lesquels j'ai travaillé par la suite et, pour les recrutements en doctorat, il m'est arrivé de faire une sélection sur un concours, que j'ai organisé à la direction régionale du Cirad. Ceci m'a permis de travailler avec de très bons étudiants. Achille Nyouma par exemple a réalisé trois articles pendant son doctorat et s'apprête à démarrer un post-doc en Espagne avec de la sélection génomique et de la GWAS chez le blé dur. Norman Munyengwa a réalisé un article pendant les 10 mois de son stage de M2 et vient de démarrer une thèse en Australie sur la sélection génomique du manguier et du merisier. Deux autres étudiants ont pour l'instant été recrutés par l'institut national sur la recherche agronomique au Cameroun, l'IRAD.

Mes affectations au **Cirad de Montpellier** ont surtout été marquées par la possibilité d'apprendre de nouveaux concepts et outils et de me former. Par exemple, depuis mon retour en 2020, j'ai découvert les principes de la structure 3D du génome et les techniques permettant de la caractériser. J'ai utilisé ces nouvelles connaissances pour écrire un projet (SelGen_3D, voir 1.6) visant à acquérir ce type de données sur quatre espèces de mon équipe, et à étudier comment les utiliser dans un contexte de prédictions génomiques. J'ai aussi créé un groupe de travail sur la simulation de schémas d'amélioration, qui me permettra à court-terme d'évoluer vers un outil de simulation plus performant (voir 3.1).

2.2. Contexte

2.2.1. Défis du monde agricole et nécessité d'une amélioration génétique plus efficace

Le monde agricole est confronté à plusieurs défis. L'augmentation régulière de la population mondiale, qui devrait atteindre 9 à 11 milliards de personnes d'ici 2050, et l'augmentation du niveau de vie dans les pays en développement génère un fort accroissement de la demande en produits dérivés des plantes, en premier lieu les produits alimentaires, mais aussi le bois, le caoutchouc naturel, etc. Ainsi, de nombreuses études prévoient une hausse autour de 80% de la demande en produits dérivés des plantes entre 2000 et 2050 (FAO, 2009 ; Le Mouël et Forslund, 2017 ; Noel et al., 2015). En parallèle, de plus de plus de contraintes pèsent sur les systèmes de culture : le changement climatique, en entrainant des accidents climatiques plus fréquents (sécheresses, inondations, etc.), des températures plus élevées, des modifications dans le régime saisonnier des précipitations, etc. ; la détérioration des sols ; la réduction de la disponibilité des ressources naturelles, et en particulier les terres arables et l'eau ; l'augmentation des stress biotiques ; et la nécessité de réduire les dégradations environnementales (pollution par les pesticides et les engrais, déforestation) (FAO, 2009 ; Noel et al., 2015 ; Tyczewska et al., 2018). Les systèmes agricoles doivent donc, à moyen-terme, produire plus malgré une pression accrue.

Parmi les solutions, il est possible d'augmenter les surfaces cultivées et le rendement des surfaces déjà cultivées. Ces deux options auront des effets environnementaux, mais l'augmentation du rendement des surfaces déjà cultivées impacte beaucoup moins les écosystèmes naturels. Depuis les années 1950, les efforts fournis dans le monde sur de nombreuses cultures ont abouti à des progrès

considérables en termes de rendement potentiel, de qualité nutritionnelle et de tolérance aux stress biotiques et abiotiques. Cependant, une augmentation des rendements d'ici 2050 selon le rythme observé depuis les années 1950 ne serait pas suffisante pour répondre à la future demande en produit dérivés des plantes (Fischer et al., 2014 ; Grafton et al., 2015 ; Tester et Langridge, 2010 ; Voss-Fels et al., 2019 ; Xu et al., 2020). Le défi actuel est donc de faire progresser le rythme d'augmentation de la productivité agricole à un niveau jamais atteint. Pour y parvenir, il est nécessaire de faire appel à de nouvelles technologies et à de nouvelles méthodes. Dans ce domaine, l'amélioration génétique a un rôle clé à jouer (Bhat et al., 2016 ; Tester et Langridge, 2010).

Beaucoup de caractères d'intérêt chez les plantes sont des caractères complexes contrôlés par un grand nombre de gènes (caractères quantitatifs), comme par exemple le rendement ou la croissance en hauteur. Les caractères quantitatifs impliqueraient des milliers de polymorphismes (Goddard et al., 2016 ; MacLeod et al., 2016). Pour ce type de caractères, l'amélioration génétique résulte classiquement d'une sélection phénotypique. Celle-ci peut se faire sur la base de la valeur propre des individus à sélectionner (sélection massale) ou de la valeur propre d'individus qui leurs sont apparentés (sélection généalogique). Le gain génétique issu de la sélection est défini comme l'amélioration de la valeur génétique moyenne d'une population sous l'effet de la sélection au cours des cycles de reproduction (Hazel et Lush, 1942). Dans le cas d'une sélection par troncation, le progrès génétique d'une génération à la suivante (ΔG) peut se prédire grâce à l'équation du sélectionneur (Walsh et Lynch, 2018, p. 490) :

$$\Delta G = r \times i \times \sigma_g$$

avec r la précision de la sélection, i l'intensité de la sélection et σ_g l'écart-type génétique. r indique la fiabilité de l'estimation de la valeur génétique des individus et se définit comme la corrélation de Pearson entre la valeur génétique réelle et la valeur génétique estimée des candidats à la sélection. i traduit la proportion d'individus sélectionnés parmi les individus évalués. σ_g représente la variabilité génétique existante au sein de la population à sélectionner (Figure 3). Le progrès génétique s'exprime souvent annuellement, en tenant compte de l'intervalle de génération, c'est à dire du nombre d'années nécessaires pour passer d'une génération à la suivante. Chez les espèces végétales, de nombreux caractères d'intérêt sont peu héritables, c-à-d fortement affectés par l'environnement. Ceci rend délicat l'estimation de la valeur génétique des individus à sélectionner. Par ailleurs, des caractères ne sont pas mesurables sur certains individus, comme les caractères de production s'exprimant uniquement chez les femelles. L'évaluation de la valeur génétique passe donc souvent par des essais spécifiques aux champs. Pour les espèces chez lesquelles il est facile de réaliser des croisements (par exemple le palmier à huile), il s'agit de tests en croisements ou tests sur descendance. Pour les espèces difficiles à croiser mais dont la multiplication végétative est facile (par exemple l'hévéa), il s'agit d'essais clonaux. Ces essais permettent ensuite de déduire la valeur génétique d'un individu à partir de la valeur propre de ses descendants ou de ses ramets. Ils donnent des estimations précises des valeurs génétiques car ils permettent de contrôler les effets environnementaux, mais en général ils sont coûteux à mettre en œuvre et augmentent la durée du cycle de sélection. Par conséquent, r est élevé mais l'intervalle de génération est grand et i est faible.

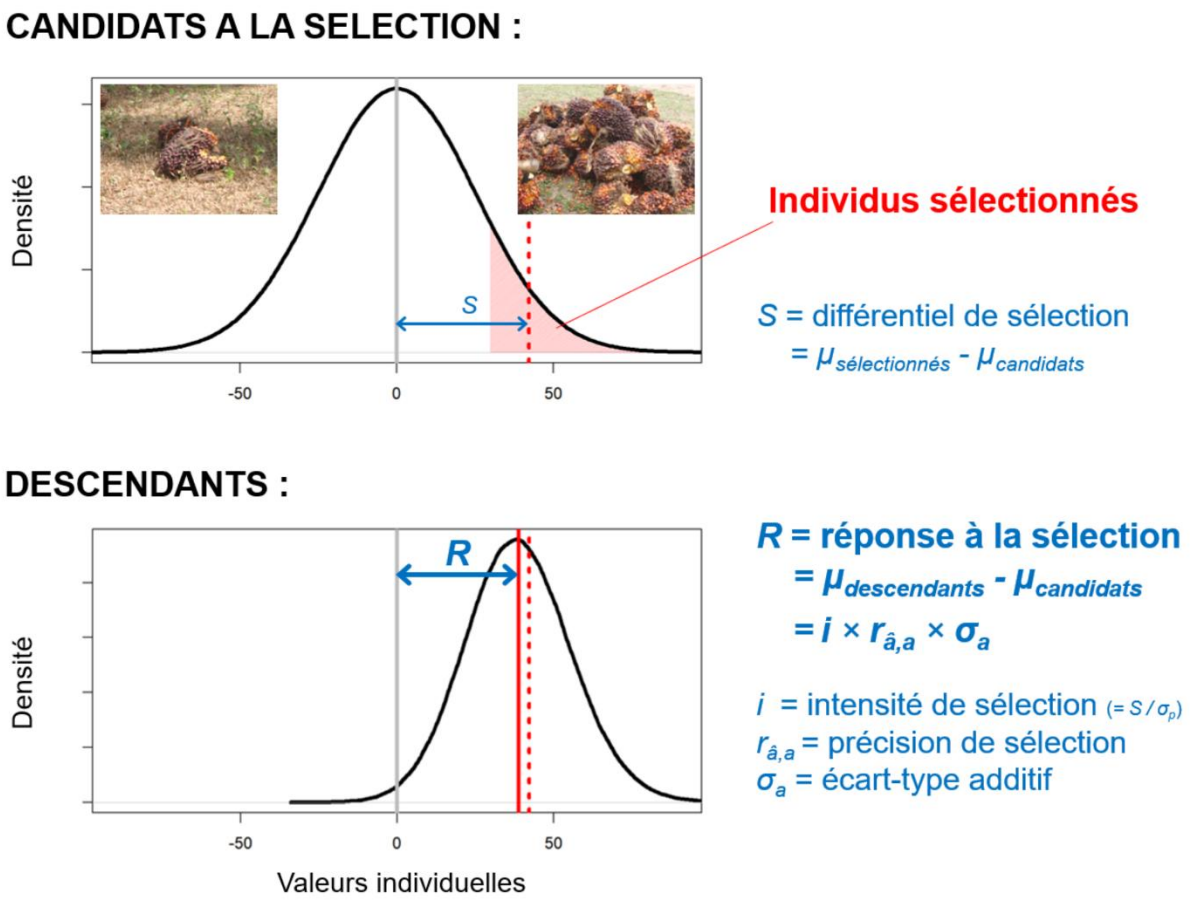


Figure 3 Réponse à la sélection et équation du sélectionneur

2.2.2. La sélection assistée par marqueurs

De nouvelles méthodes, et en particulier des méthodes de sélection assistée par marqueurs (SAM) efficaces pour les caractères quantitatifs sont nécessaires pour contourner les limites des approches classiques purement phénotypiques.

La SAM pour les caractères quantitatifs fait l'objet de beaucoup d'intérêt depuis plusieurs décennies. Des approches de SAM reposent sur la détection de zones du génome contrôlant le caractère d'intérêt (QTL). Elles peuvent largement augmenter la vitesse, l'efficacité et la précision de l'amélioration génétique comparé aux approches phénotypiques (Gupta et al., 2010). Lande et Thompson (1990) ont montré que la précision de ce type de SAM pour des caractères faiblement héritables pouvait en théorie largement dépasser la précision de la sélection phénotypique. Cette perspective très prometteuse n'a cependant pas eu les retombées espérées. En effet, cette approche est peu efficace pour détecter les QTL à effets faibles. En identifiant uniquement les QTL à effets forts, le sélectionneur n'a accès qu'à une part modeste des effets génétiques contrôlant véritablement les caractères complexes. De plus, pour des raisons liées à la méthode statistique, les effets des QTL mis en évidence étaient souvent surestimés.

Finalement, la SAM basée sur la détection de QTL est efficace uniquement pour les caractères contrôlés par un petit nombre de QTL ; pour les caractères quantitatifs, elle peut s'avérer moins efficace que la sélection phénotypique. Chez les espèces forestières, les plantes cultivées et les animaux, plusieurs décennies d'efforts de recherche en matière de SAM basée sur des approches de

détection de QTL n'ont ainsi pas donné de résultats probants en pratique (Grattapaglia et al., 2018, p. 2 ; Muranty et al., 2014).

2.2.3. La sélection génomique

a. Origine du concept

La sélection génomique (SG) s'appuie sur le modèle linéaire mixte avec, souvent, une analyse par la méthodologie BLUP (*best linear unbiased predictor*) (Henderson, 1950 ; Schaeffer, 1991). Leurs principes généraux, qui ne sont pas spécifiques à la SG, sont détaillés dans l'Annexe 3.

En 1994, Bernardo a développé une méthode pour prédire la valeur de croisements hybrides de maïs non phénotypés (Bernardo, 1994). Elle se basait sur le modèle linéaire mixte et le BLUP mais en remplaçant la traditionnelle matrice généalogique des apparentements entre individus par une matrice d'apparentements moléculaires, calculée à partir de 220 marqueurs de type RFLPs (*restriction fragment length polymorphism*) selon la méthode décrite par Lynch (1988). Le calcul des apparentements à partir de marqueurs moléculaires présente plusieurs avantages par rapport à un calcul à partir du pédigrée. Les apparentements généalogiques sont des apparentements attendus, qui peuvent dévier des apparentements réalisés pour plusieurs raisons : le pédigrée considère que les individus fondateurs ne sont pas apparentés, l'échantillonnage mendélien (c-à-d la ségrégation au sein des familles de plein-frères) et la sélection sont négligés, et le pédigrée peut contenir des erreurs (individus illégitimes). L'utilisation de marqueurs permet de ne pas être confronté à ces problèmes. Cependant, le potentiel de la méthode n'a pas vraiment été mesuré à l'époque, l'article de Bernardo ayant eu relativement peu de retombées par rapport à celui publié en 2001 par Meuwissen et al. (Figure 4).

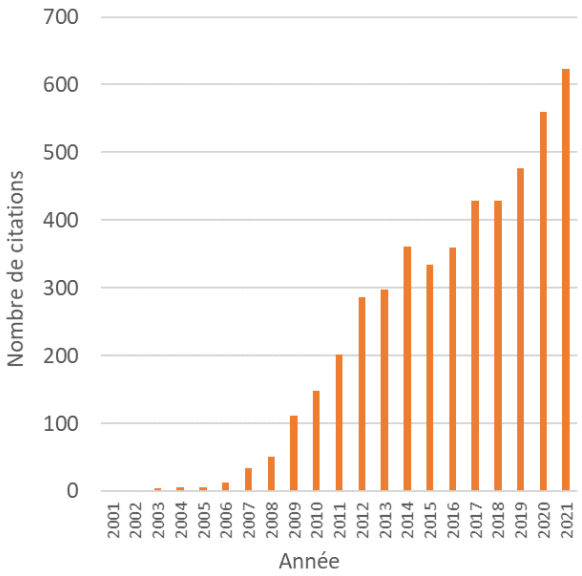


Figure 4 Nombre de citations annuelles de l'article Meuwissen et al 2001 (Genetics) d'après Google Scholar

Meuwissen et al. (2001) ont proposé, dans une étude par simulation, le concept de SG pour prédire, à partir d'un modèle linéaire mixte et d'un marquage dense sur tout le génome, la valeur génétique d'animaux non phénotypés. Cet article introduisait trois méthodes statistiques pour les prédictions, basées sur le BLUP (méthode plus tard nommée RRBLUP, pour *random-regression* ou *ridge-regression* BLUP) et sur des approches Bayésiennes (BayesA et BayesB). La SG suppose que la densité de marquage est suffisante pour avoir chaque QTL en déséquilibre de liaison avec au moins un marqueur. Les méthodes de prédiction proposées estimaient des effets aux marqueurs utilisés pour déduire la valeur génétique des individus, contrairement à l'approche de Bernardo qui fournissait directement une prédiction des valeurs génétiques. Une autre approche de SG, similaire à celle de Bernardo, a par la suite été suggérée, le GBLUP (pour *genomic* BLUP) (VanRaden, 2008 ; VanRaden, 2007). Le GBLUP et le RRBLUP sont équivalents lorsqu'un

grand nombre de QTL sont impliqués, qu'il n'y a pas QTL majeurs et que les QTL sont régulièrement répartis le long du génome (Bernardo, 2020 ; Habier et al., 2007).

En 2001, les méthodes de génotypage haut-débit nécessaires pour la mise en pratique de la SG n'étaient pas disponibles. Aucun génome complet n'avait même été publié chez les animaux d'élevage et les plantes cultivées (Feuillet et al., 2011 ; de Koning, 2016). La mise en œuvre de la SG a été rendue possible par le développement des technologies de séquençage de nouvelle génération (NGS, pour *next generation sequencing*), devenues accessibles entre 2004 et 2006 (Hu et al., 2021). Moins coûteuses et d'un débit très largement supérieur par rapport à la méthode Sanger (Sanger et Coulson, 1975 ; Sanger et al., 1977), les méthodes de NGS ont permis de réaliser à un coût abordable du génotypage haute-densité et haut-débit, c-à-d avec une bonne couverture du génome sur de grandes populations.

Une fois ces conditions réunies, et compte tenu des résultats prometteurs obtenus dans les premières études, l'intérêt pour la SG a augmenté très rapidement chez les animaux et les plantes (Figure 4).

b. Principe

La première étape de la SG consiste à créer une population de calibration (ou d'entraînement) composée d'individus pour lesquels on dispose du génotype et d'observations (valeur propre ou estimation de la valeur génétique) pour les caractères cibles. Dans la seconde étape, un modèle de prédiction est construit à l'aide de ces données et appliqué à la population de sélection, génotypée avec les mêmes marqueurs mais pour laquelle aucune observation phénotypique n'est disponible. Le modèle prédit la valeur génétique des candidats à la sélection, qui correspond, selon le modèle, à la valeur génétique additive (GEBV, *genomic estimated breeding values*) ou totale (Grattapaglia et al., 2018 ; Heffner et al., 2009) (Figure 5).

La précision de la SG est définie comme la corrélation de Pearson entre la valeur génétique prédite et la valeur génétique des candidats à la sélection. Cependant, dans les études empiriques, la valeur génétique des individus de validation n'est pas connue. La corrélation calculée est alors la précision de prédiction (*predictive ability* ou *prediction accuracy*). Elle mesure la capacité du modèle de SG à prédire les valeurs observées, et non la valeur génétique. Si les valeurs observées sont des phénotypes, on peut en déduire la précision de la SG. Pour la prédiction des valeurs additives, elle vaut $r_{GEBV,A} = r_{GEBV,P} / h$, en supposant que les erreurs associées aux GEBV et aux phénotypes soient indépendantes (Legarra et al., 2008, p. 618 ; Lorenz et al., 2011, p. 94). La précision de la SG est généralement obtenue par validation croisée en *k* parties (*k-fold cross-validation*) au sein d'un même dispositif expérimental, chaque partie étant utilisée alternativement comme population de validation et les parties restantes comme population de calibration ; ou entre dispositifs expérimentaux, un site étant utilisé pour la calibration et l'autre pour la validation. Cette dernière méthode est préférable car les validations croisées peuvent surestimer la précision (Lorenz et al., 2011).

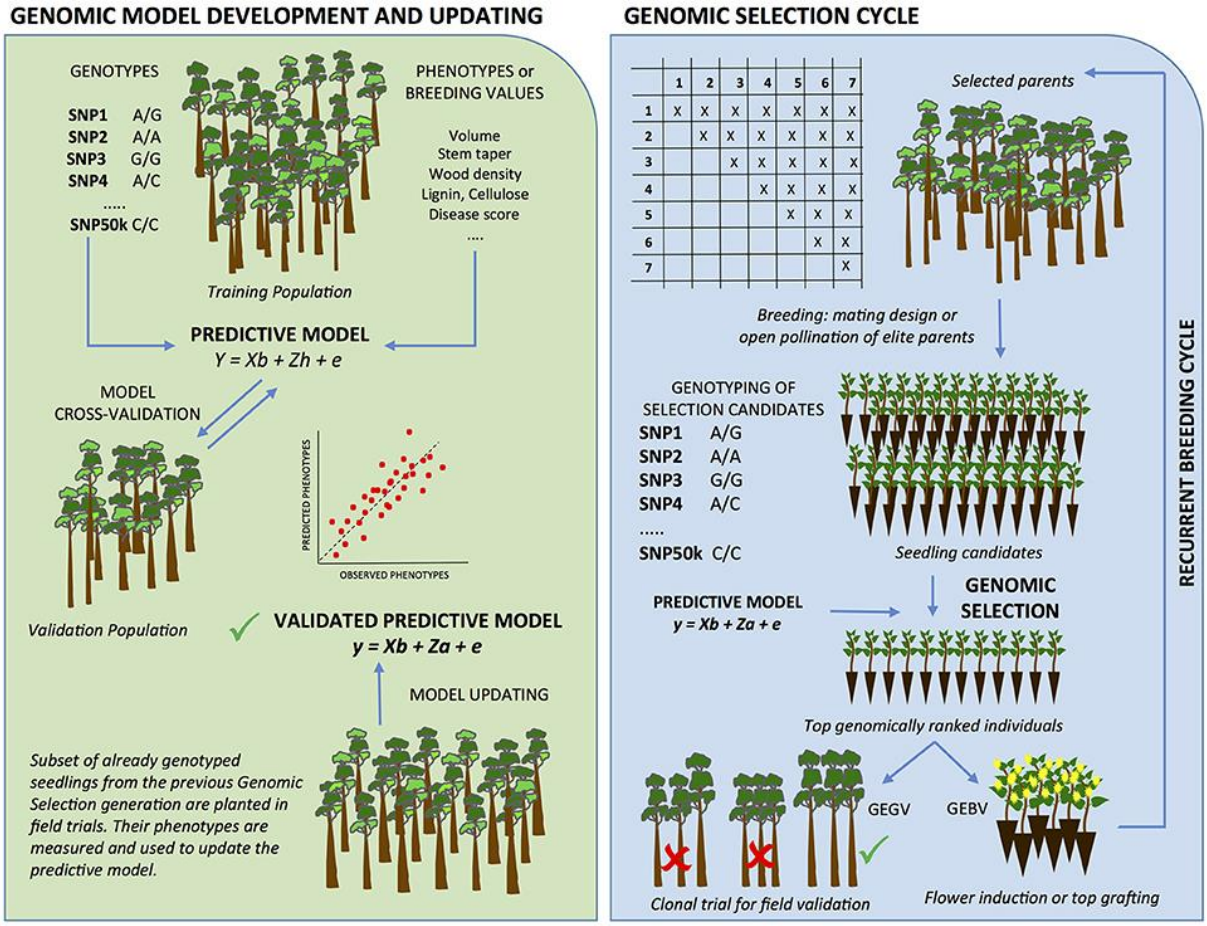


FIGURE 1 | Genomic selection in forest trees. GS begins with the development of a predictive model for the traits of interest **(Left panel)**, which are then used in the GS cycles **(Right panel)** and progressively updated. GS uses genome-wide markers whose effects on the phenotype are estimated concurrently in a large and representative "training population" of individuals without applying severe significance tests. Markers are retained as forecasters of phenotypes in prediction models to be later applied to "selection candidates" for which only genotypes are collected. The prediction models are cross-validated against a "validation population," a set of individuals of the same reference population that were not used for the estimation of marker effects. Once a prediction model is shown to provide adequate accuracy, it can be used in the GS cycle. An array of selection candidates - full of half-sib families derived from crossing either the original elite parents of the training set, or elite individuals selected in the training set - are genotyped and have their breeding values (GEBV) and/or genotypic values (GEGV; additive + non-additive effects) estimated using the model developed earlier. Top ranked seedlings for GEBV are subject to early flower induction and inter-mated to create the next generation of breeding. Top ranked seedlings for GEGV are clonally propagated and tested in verification clonal trials where elite clones are eventually selected for operational plantation. Additionally, all or subsets of the already genotyped selection candidates are planted in experimental design and phenotyped at the target selection age to provide genotype and trait data for GS model updating as GS generations advance and climate changes.

Figure 5 Schéma de principe de la sélection génomique (Grattapaglia et al., 2018)

c. Méthodes statistiques de prédiction

Dans les approches de détection de QTL, tous les marqueurs ne sont pas considérés simultanément dans le modèle et un test de signification est réalisé pour conserver ou rejeter les marqueurs. En conséquence, l'effet des QTL correctement identifiés et la variance phénotypique expliquée par ces QTL sont surestimés, en particulier lorsque le nombre d'individus phénotypés est petit (exemple 100) (Beavis, 1994 ; Beavis, 1998). Ceci rend quasiment impossible la détection des QTL à faibles effets.

Les modèles de régression sur l'ensemble du génome utilisés pour les prédictions génomiques traitent du problème « grand nombre de variables, petit nombre de données » qui, dans la SG, concerne le nombre de marqueurs qui dépasse généralement (largement) le nombre d'observations phénotypiques. Les modèles de SG permettent de considérer conjointement l'ensemble des marqueurs, ce qui permet de leur estimer des effets associés non biaisés et de prendre en compte les

QTL ayant de petits effets. Les marqueurs sont ensuite tous utilisés, sans test de signification de leur effet.

Un large éventail de méthodes statistiques a été développé pour la SG afin de pallier cette contrainte (Figure 6) (de los Campos et al., 2013 ; Garrick et al., 2014 ; Montesinos López et al., 2022 ; Morota et Gianola, 2014 ; Tong et Nikoloski, 2021 ; Wang et al., 2018). Elles se répartissent en deux grandes catégories :

(i) les approches paramétriques, qui comprennent principalement les méthodes qui s'appuient sur la méthodologie BLUP, c-à-d le GBLUP (VanRaden, 2008 ; VanRaden, 2007) et le RRBLUP (Meuwissen et al., 2001), et diverses méthodes bayésiennes, comme la *Bayesian ridge regression* (Pérez et al., 2010), le LASSO Bayésien (de los Campos et al., 2009), BayesA, BayesB (Meuwissen et al., 2001), BayesCπ et Bayes Dπ (Habier et al., 2011),

(ii) les approches semi- et non paramétriques, qui entrent dans la catégorie du *machine learning* et qui comprennent, en particulier, le *reproducing kernel Hilbert spaces* (RKHS) (Gianola et van Kaam, 2008) et les réseaux de neurones artificiels (Montesinos-López et al., 2021 ; Tong et Nikoloski, 2021).

Ces méthodes diffèrent à plusieurs égards : en termes d'hypothèses génétiques et de modélisation de l'architecture génétique des caractères avec, par exemple, des modèles purement additifs ou qui modélisent explicitement les effets de dominance et/ou épistatiques, des modèles avec des effets de marqueurs échantillonnés à partir d'une distribution statistique commune (RRBLUP, GBLUP, *Bayesian Ridge Regression*) ou à partir de distributions spécifiques (LASSO Bayésien, BayesB, etc.), en termes d'approche de calcul (méthodes basées sur les apparentements génomiques et celles estimant des effets aux marqueurs, modèles uni-variés et multi-variés, etc.), et en termes d'informations génomiques utilisées dans le modèle (type de polymorphismes, utilisation d'informations *a priori* sur les marqueurs, combinaison de données -omiques, etc.)

La méthode statistique la plus couramment utilisée en SG est le GBLUP (Heslot et al., 2015 ; Montesinos-López et al., 2021). Le modèle de base est de la forme :

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e} \quad \text{(Équation 1)}$$

avec $\boldsymbol{Y}$ le vecteur des *n* observations de la population de calibration (dimension *n*×1), $\boldsymbol{\beta}$ le vecteur des effets fixes (*p*×1), $\boldsymbol{u}$ le vecteur des valeurs génétiques additives des *q* individus considérés (*q*x1), $\boldsymbol{X}$ la matrice d'incidence associant les observations aux effets fixes (*n*x*p*), $\boldsymbol{Z}$ la matrice d'incidence associant les observations aux valeurs génétiques additives (*n*x*q*) et $\boldsymbol{e}$ le vecteur des effets résiduels (*n*×1), avec $\boldsymbol{e} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I\sigma_e^2})$ et $\boldsymbol{I}$ une matrice identité. $\boldsymbol{u}$ est un effet aléatoire associé à une matrice d'apparentements génomiques $\boldsymbol{G}$ incluant les individus des populations de calibration et d'application, et telle que $\boldsymbol{u} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{V_u})$, avec $\boldsymbol{V_u} = \boldsymbol{G\sigma_a^2}$ et $\boldsymbol{\sigma_a^2}$ la variance additive (VanRaden, 2007).

Pour le RRBLUP, le modèle de base est de la forme (Meuwissen et al., 2001 ; Mrode, 2014, p. 183):

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z'm} + \boldsymbol{e}$$

où $\boldsymbol{m}$ est le vecteur des effets de substitution à estimer pour les *k* marqueurs utilisés (*k*×1), considéré comme un effet aléatoire et tel que $\boldsymbol{m} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{I\sigma_m^2})$, et $\boldsymbol{Z'}$ la matrice d'incidence donnant le génotype des individus de la population de calibration aux *k* marqueurs. Le vecteur $\widehat{\boldsymbol{u}}^*$ des GEBV des individus de la population d'application s'obtient par : $\widehat{\boldsymbol{u}}^* = \boldsymbol{Z}^*\widehat{\boldsymbol{m}}$, avec $\widehat{\boldsymbol{m}}$ le vecteur des effets aux marqueurs estimés dans la population de calibration et $\boldsymbol{Z}^*$ la matrice des génotypes des individus de la population d'application.

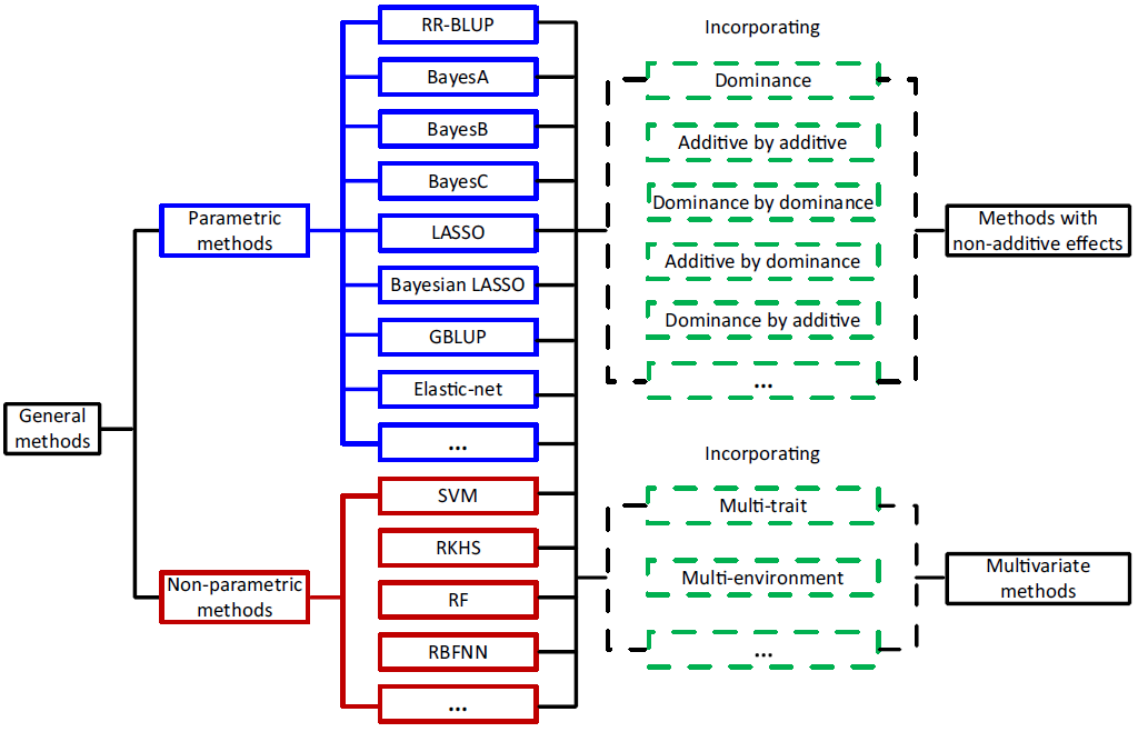


Figure 6 Vue d'ensemble des principales approches statistique pour les prédictions génomiques (Wang et al., 2018)

d. Génotypage

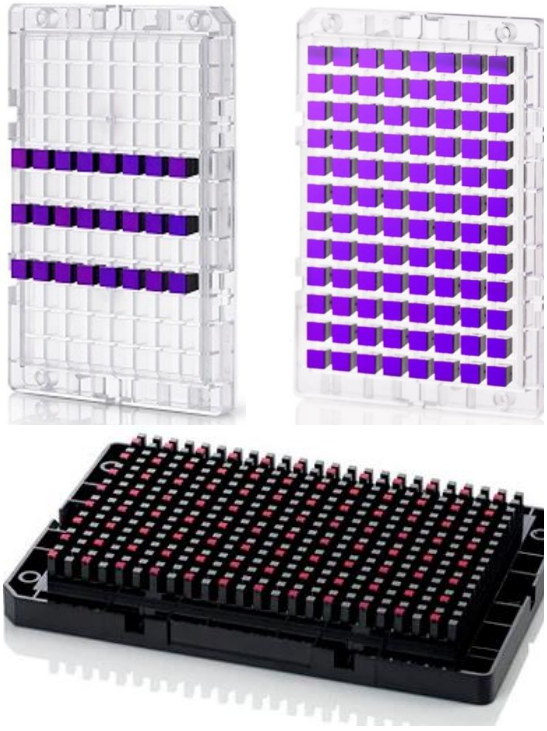


Figure 7 Plaques de génotypage Axiom™ pour 24, 96 et 384 échantillons, avec des puces pouvant contenir plusieurs millions de SNP

Les premières technologies NGS sur lesquelles s'est appuyée la SG, telles que Ion Torrent, 454 et Illumina/Solexa, sont aujourd'hui dites de seconde génération (la première génération correspondant à la méthode Sanger). Elles produisent des *short-reads*, c-à-d des séquences de fragments de tailles inferieures à 1 kb (van Dijk et al., 2014 ; Hu et al., 2021 ; Kumar et al., 2019). Avec elles, deux options sont devenues techniquement et économiquement envisageables pour acquérir les données moléculaires nécessaires à la SG : les puces à SNP et les méthodes de réduction de la complexité du génome (Edwards et al., 2013 ; Wiggans et al., 2017). Le développement des puces à SNP (voir illustrations Figure 7) demande des efforts en termes de séquençage de génomes entiers et de reséquençage afin d'identifier des SNP valides. Les puces à SNP sont ainsi obtenues à partir du séquençage de profondeur réduite (reséquençage) d'un échantillon d'individus représentatifs de la diversité et dont les *reads* seront assemblés grâce à la séquence de référence du génome de l'espèce. Le polymorphisme est mis en évidence en comparant la séquence des individus utilisés. De nombreux exemples de développement de puces à SNP ont été publiés chez les plantes, comme par exemple

chez l'épicéa (Bernhardsson et al., 2021), le pommier (Chagné et al., 2012) et le palmier à huile (Kwong et al., 2016). Différentes méthodes basées sur la réduction de la complexité du génome existent (Edwards et al., 2013 ; Ray et Satya, 2014 ; Zhou et Holliday, 2012). Il s'agit notamment des méthodes basées sur l'utilisation d'enzymes de restriction, et en particulier le génotypage par séquençage (GBS, *genotyping-by-sequencing*) (Elshire et al., 2011), et les méthodes de *sequence capture* (Zhou et Holliday, 2012). Ces méthodes associent la découverte de marqueurs et le génotypage, sans nécessiter d'étude préalable de polymorphisme. Elles ne demandent donc pas d'investissement préliminaire et peuvent être appliquées directement sur n'importe quelle population, mais présentent un taux plus élevé de données manquantes et d'erreurs de génotypage que les puces à SNP. Le GBS, par exemple, repose sur le séquençage des régions délimitées par le site de restriction des enzymes utilisées, et peut cibler préférentiellement les zones du génome riches en gènes (zones avec des séquences à faible fréquence) et éliminer les zones répétées, grâce à une combinaison adaptée d'enzymes de restriction (Figure 8).

Aujourd'hui, les technologies de séquençage de troisième génération, telles que le *single-molecule real-time sequencing* de Pacific Biosciences et le *nanopore sequencing* de Oxford Nanopore Technologies permettent de séquencer des fragments entre 1 kb et 4 Mbp (*long-reads sequencing*). Les méthodes les plus performantes (PacBio HiFi) présentent des taux d'erreurs de séquençage similaires à ceux des méthodes de seconde génération. Les technologies de séquençage de troisième génération permettent de mieux assembler et phaser les données (Figure 9), ainsi que de détecter les variants structuraux (réarrangement génomiques de plus de 50 bp) (De Coster et al., 2021 ; Hu et al., 2021 ; Kumar et al., 2019). Ces nouvelles approches devraient contribuer à améliorer les performances de la SG (voir 3.4.1 et 3.4.2).

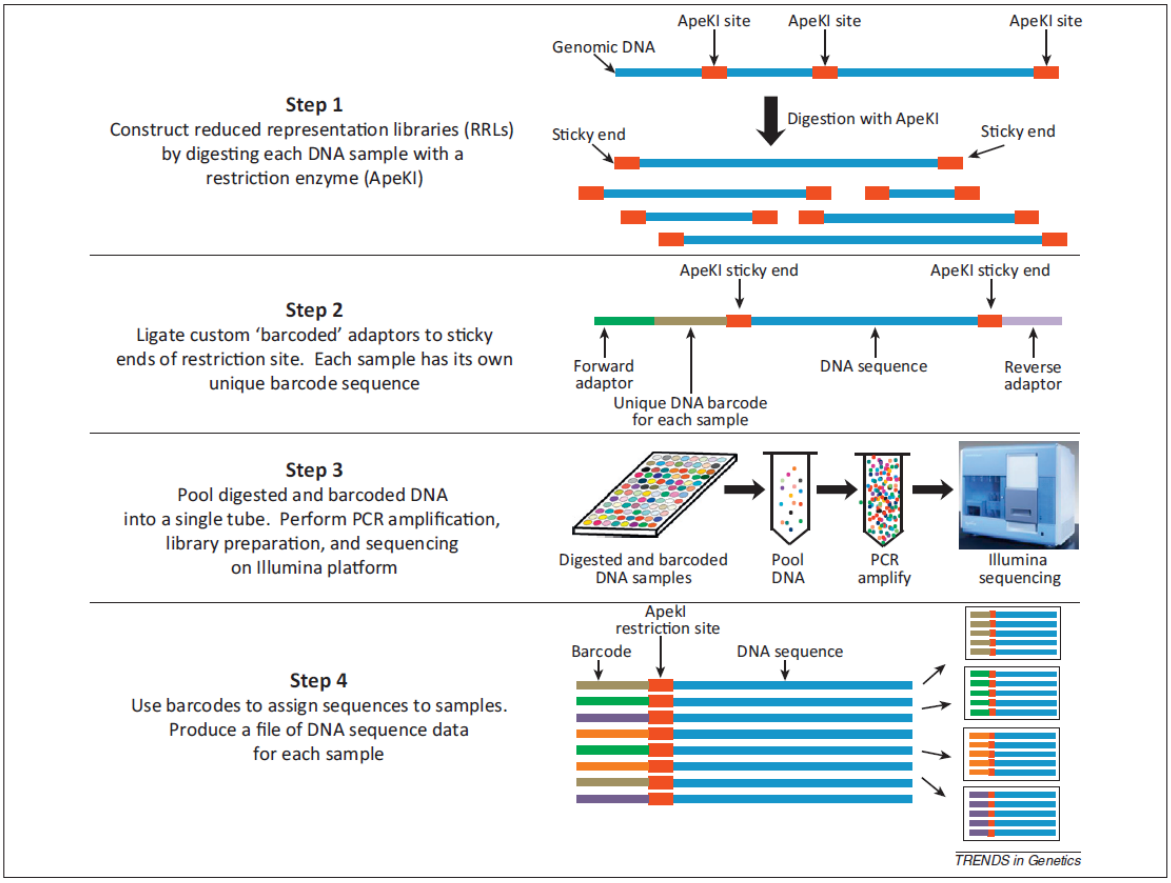


Figure 8 Le principe du génotypage par séquençage (GBS, genotyping by sequencing) (Myles, 2013)

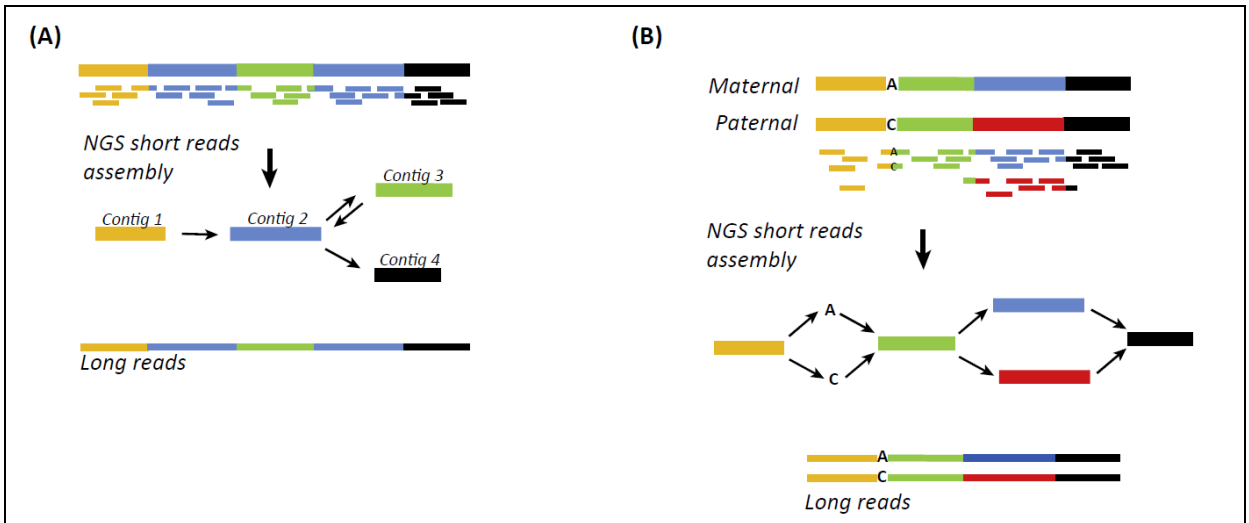


Figure 1. (A) Example showing the resolution of a repeated genome region by short-read assembly or long reads. Sequencing a region with two nearly identical repeats (blue) separated by a unique sequence will generate reads corresponding to the upstream region (yellow), the repeats, the sequence between (green), and the downstream region (black), and some reads will overlap the boundaries. Assembly programs cannot assign reads falling in the repeats to unique positions and will assemble those reads into a single contig. The sequence between the repeats cannot be assigned to a unique position either as it can be placed either upstream or downstream of the 'blue' region. Due to this ambiguity, the sequences upstream of, between, and downstream of the repeats will be assembled into separate contigs. Similar problems arise with structural variants that involve repetitive regions. (B) Haplotype phasing. SNPs (single nucleotide: A or C) or larger variations (red or blue) between maternal and paternal alleles located too far apart to be covered by a single read will be difficult to phase to the parental allele of origin. This will lead to ambiguous trajectories that result in fragmented assemblies.

Figure 9 Comparaison des performances des méthodes de séquençage de seconde génération (*short-reads)* et de troisième génération (*long-reads*) en termes d'assemblage (A) et de phasage (B) (van Dijk et al., 2018)

e. Applications pratiques et retombées

La SG est devenue une des approches les plus prometteuses pour améliorer le gain génétique annuel et/ou par unité de coût chez les animaux et les plantes (Fugeray-Scarbel et al., 2021 ; Mrode et al., 2019 ; Voss-Fels et al., 2019 ; Wartha et Lorenz, 2021 ; Xu et al., 2020). Chez les bovins laitiers, la SG a doublé le rythme du progrès génétique (Wiggans et al., 2017). Chez les plantes, elle est progressivement intégrée aux schémas d'amélioration (Merrick et al., 2022 ; Varshney et al., 2017 ; Voss-Fels et al., 2019).

Les premières études sur la SG visaient essentiellement à estimer la précision des prédictions. Cependant, bien que cette précision soit d'un intérêt majeur pour le sélectionneur, elle n'est pas suffisante pour mesurer l'efficacité d'un schéma de sélection, qui dépend aussi d'autres facteurs, selon l'équation du sélectionneur (voir 2.2.1). Ainsi, par exemple, si la SG est utilisée pour amener un accroissement important de l'intensité de la sélection, elle augmentera le gain génétique même si sa précision est un peu moins élevée qu'avec l'évaluation phénotypique conventionnelle. Les études qui ont suivi ont donc souvent étendu les analyses jusqu'à l'estimation du progrès génétique annuel, qui est le véritable paramètre permettant d'identifier le meilleur schéma d'amélioration parmi différentes options, et en particulier des alternatives incluant des prédictions génomiques. Ces études-ci ont montré que, en fonction de la biologie de l'espèce considérée et de son schéma d'amélioration, la sélection génomique peut permettre d'accroître le progrès génétique annuel par rapport à la sélection classique de plusieurs façons (Fugeray-Scarbel et al., 2021 ; Wartha et Lorenz, 2021) :

(i) en augmentant l'intensité de sélection, lorsque le facteur limitant en sélection classique est le nombre d'individus que l'on peut phénotyper,
(ii) en raccourcissant l'intervalle de génération, en remplaçant des étapes de phénotypage par du génotypage, plus rapide,
(iii) en augmentant la précision de sélection, en particulier pour les caractères difficiles à phénotyper.

Le potentiel de la sélection génomique est donc particulièrement élevé pour les plantes pérennes et encombrantes, car elles ont souvent (Isik, 2014):

- un grand intervalle de génération (>10 ans), à cause d'une expression tardive des caractères d'intérêt et/ou du besoin d'un phénotypage dans des essais spécifiques nécessitant plusieurs années (essais clonaux ou tests sur descendance)
- une intensité de sélection contrainte à cause du coût élevé et de la complexité de dispositifs expérimentaux couvrant des superficies importantes, limitant le nombre d'individus testés (population de sélection <500 individus).

2.2.4. Le palmier à huile et son amélioration génétique

a. La plante

Le palmier à huile (*Elaeis guineensis* Jacquin) est une monocotylédone pérenne de la famille des Arécacées (Corley et Tinker, 2016 ; Soh et al., 2017). Il est originaire d'Afrique et son aire naturelle s'étend sur plus de 6 000 km le long de la côte Atlantique d'Afrique depuis le Sénégal jusqu'à l'Angola, et s'enfonce sur 50 à 200 km à l'intérieur des terres, et sur 2 000 km au niveau de l'équateur, dans la cuvette congolaise.

Figure 10 Palmier à huile en plantation

Le palmier à huile est une herbe géante qui produit tout au long de l'année des feuilles, entourant le bourgeon végétatif pour former la couronne. Les feuilles mesurent 6 à 9 mètres et sont composées de plus de 300 folioles. A l'aisselle de chaque feuille se trouve une inflorescence dont le devenir (sexualisation femelle ou mâle, ou avortement) dépend des conditions environnementales au cours de son développement, en particulier du bilan hydrique, et des cycles sexuels endogènes du palmier. Une fois fécondées, les inflorescences femelles évoluent normalement en régimes (Figure 10). Des illustrations sur la plante et sa filière sont fournies en Annexe 4.

Un régime est constitué d'un rachis (ou pédoncule) portant des épillets, sur lesquels se trouvent les drupes (fruits à noyaux). Un régime pèse entre 5 et 50 kg, selon l'âge du palmier, sa population d'origine, son environnement, etc. Un fruit se compose généralement d'une amande (faite

d'un embryon et d'albumen), d'un endocarpe ligneux (coque), de mésocarpe (pulpe) et d'un exocarpe (peau).

Chez le palmier à huile coexistent trois types, définis par la morphologie interne de leurs fruits :

- le dura : il s'agit du type prépondérant dans la nature (>90%). Ses fruits possèdent une coque épaisse (de 2,5 à 7 mm) et un pourcentage de pulpe assez faible.
- le pisifera : il est très rare dans la nature (<5%). Ses fruits sont dépourvus de coque et sa pulpe renferme des fibres lignifiées qui, lors d'une coupe transversale du fruit, forment un anneau autour de l'amande. Les pisifera sont généralement improductifs car leurs régimes avortent avant maturité. Les fruits de pisifera sont donc très rares mais lorsqu'ils existent ils possèdent un pourcentage de pulpe très élevé.
- le tenera : il est très rare dans la nature (<5%). Ses fruits possèdent une coque de faible épaisseur (<2 mm), et un anneau de fibres lignifiées dans la pulpe, autour du noyau.

Le déterminisme génétique de la présence ou de l'absence de coque a été mis en évidence dans les années 1930 (Beirnaert et Vanderweyen, 1941). Ce caractère est sous le contrôle d'un gène nommé *Sh* (*shell*). Du point de vue statistique, celui-ci possède deux allèles codominants, *Sh+* qui permet la formation d'une coque et un mutant d'effet opposé *Sh-*. Les dura sont donc de génotype *Sh+//Sh+* et les pisifera *Sh-//Sh-*. Leur hybride, le tenera, est hétérozygote et présente un phénotype intermédiaire. Du point de vue biologique, le *gène Shell* code pour un facteur de transcription de la famille des MADS-box, et l'allèle *Sh-* correspond en fait à dix mutations (neuf faux-sens et une délétion). Les protéines de type MADS-box forment normalement des dimères, et les mutations Sh- perturbent la dimérisation et/ou la fixation du dimère sur l'ADN (Ooi et al., 2016 ; Singh et al., 2013a ; Singh et al., 2020).

Le palmier à huile commence à produire des inflorescences vers trois ans. Bien que monoïque, le palmier à huile a une reproduction rendue allogame par l'alternance des cycles mâles et femelles (dioécie temporelle). La pollinisation est entomophile, et implique en particulier des charançons du genre *Elaeidobius* (Li et al., 2019). Il n'y a pas de reproduction végétative naturelle mais elle est possible, bien que délicate, par culture *in vitro*. Le clonage du palmier à huile a été ralenti par l'apparition d'une morphogenèse florale anormale au champ (variants *mantled*), conduisant à des palmiers stériles (Soh et al., 2017, p. 172). Le mécanisme moléculaire épigénétique à l'origine de cette anomalie a été récemment élucidé (Ong-Abdullah et al., 2015 ; Soh et al., 2017, p. 207), relançant l'intérêt pour le clonage.

Le palmier à huile est diploïde et possède 16 paires de chromosomes (2n=32). Son génome couvre une distance génétique d'environ 1 500 cM (Seyum et al., 2022b ; Yue et al., 2021) et 1,8 Gb (Singh et al., 2013b). Plusieurs séquences du génome nucléaire ont été acquises par des consortiums internationaux ou de grandes entreprises (Jin et al., 2016 ; Murphy, 2014 ; Singh et al., 2013b ; Wang et al., 2022). Le *Malaysian Palm Oil Board* a rendu publique une séquence complète sous forme de 16 pseudo-molécules chromosomiques (Singh et al., 2013b). Une version améliorée par Sime Darby et couvrant 1,2 Gb est elle aussi publique (Ong et al., 2020). L'utilisation du séquençage de troisième génération a aussi permis d'obtenir, pour un autre individu, un génome assemblé avec 16 pseudo-chromosomes couvrant 1,56 Gb (Wang et al., 2022). Environ 35 000 gènes ont été prédits par similarité avec des protéines connues (Singh et al., 2013b). La comparaison des chromosomes a révélé l'existence de nombreuses régions dupliquées dans le génome.

Les principaux stress biotiques auxquels est soumis le palmier à huile sont des maladies, qui peuvent avoir une incidence économique très forte, avec une mortalité importante en plantation. La fusariose vasculaire est causée par un champignon du sol, *Fusarium oxysporum* f. sp. *Elaeidis*. Elle se rencontre en Afrique. La maladie du ganoderma, ou pourriture basale du stipe, est causée par un autre

champignon du sol, *Ganoderma boninense*, sévissant en Asie et, de plus en plus, en Afrique. Enfin, en Amérique Latine, les palmiers à huile sont attaqués par la pourriture du cœur, dont l'agent pathogène n'a pas encore été identifié.

Le genre *Elaeis* compte une autre espèce, *E. oleifera* originaire d'Amérique latine. Les deux espèces auraient divergé il y a environ 51 millions d'années (Singh et al., 2013b). L'espèce *E. oleifera* se distingue par une forte teneur en acides gras insaturés, une croissance en hauteur lente et une résistance à certains parasites et ravageurs (pourriture du cœur, mineuse des feuilles), mais présente un rendement extrêmement faible en huile. L'utilisation commerciale d'*E. oleifera* se limite à la production d'hybrides interspécifiques pour des zones où la pourriture du cœur empêche la culture de l'espèce africaine. Dans ce document, il est uniquement question de l'espèce *E. guineensis*.

b. Huile de palme et huile de palmiste

L'huile de palme, tirée de la pulpe des fruits, et l'huile de palmiste, tirée de l'amande, font parties des graisses concrètes, c-à-d des huiles solides à température ambiante (Lecerf, 2017).

L'huile de palme contient près de 100 % de lipides, sous forme de triglycérides, constitués d'une molécule de glycérol associée à trois acides gras (Figure 11). Elle est composée de 45 à 55 % d'acides gras saturés, majoritairement de l'acide palmitique C16:0 (39 à 47 % des acides gras), et 45 à 55 % d'acide gras insaturés, majoritairement de l'acide oléique C18:1 (36 à 44 % des acides gras), le reste étant de l'acide linoléique C18:2 (9-12 % des acides gras). L'huile de palme brute est l'huile la plus riche en caroténoïdes (20 fois plus riche que l'huile d'olive et 200 fois plus que l'huile de tournesol) et en tocotriénols, des composés bénéfiques pour la santé et à l'origine de sa couleur rouge (Annexe 4 L). L'huile de palme consommée dans les pays du Nord est essentiellement sous sa forme raffinée, très largement appauvrie en caroténoïdes. Sa mauvaise réputation dans les pays du Nord en termes d'effets sur la santé apparait actuellement injustifiée d'un point de vue scientifique (Lecerf, 2017 ; Rival et Levang, 2013). Aucune huile n'est parfaite du point de vue de sa composition, et la consommation d'huile de palme au sein d'un régime alimentaire équilibré et varié ne présente pas de risque pour la santé (Absalome et al., 2020)

L'huile de palmiste a une composition chimique semblable à l'huile de coprah (tirée de la noix de coco), avec une très forte proportion en acides gras saturés (>80%).

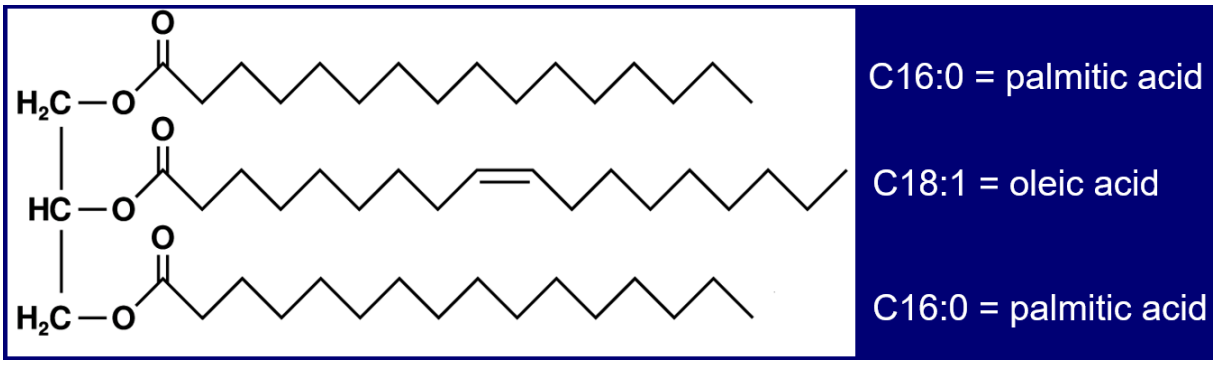


Figure 11 Exemple d'un triglycéride

c. Utilisations

L'huile de palme sert à 80% dans l'alimentation humaine (huile de table, huile de friture, margarine, etc.), mais elle a aussi des débouchés dans l'industrie, avec 19% des usages en oléochimie (cosmétique, savonnerie, lubrifiants, etc.) et 1% dans le biodiesel (Rival et Levang, 2013, p. 18).

Dans certains pays, en particulier en Afrique, l'huile de palme est la principale source de corps gras dans le régime alimentaire. Elle joue alors un rôle majeur dans les apports lipidiques, énergétiques et vitaminiques. En France, la consommation moyenne d'huile de palme par habitant est faible et se situe autour de 2 kg / personne / an (Rival et Levang, 2013, p. 53), soit environ 6% de la consommation totale de lipides des adultes.

L'huile de palmiste sert dans l'alimentation humaine (margarine par exemple) mais aussi en savonnerie, cosmétique et oléochimie.

d. Filière et production

Le palmier à huile commence à produire lors de sa 3ème ou 4ème année, selon l'environnement, et est en général exploité pendant une vingtaine d'années, la hauteur des plantes rendant alors la récolte difficile. La densité de plantation est normalement de 143 palmiers par hectare. Une des principales qualités du palmier à huile parmi les autres plantes oléagineuses est sa productivité, qui atteint 3,8 tonnes d'huile de palme par hectare en moyenne mondiale, loin devant l'ensemble des autres oléagineuses (Rival et Levang, 2013). Ce rendement exceptionnel, associé à un faible coût de production, a largement contribué à ce que le palmier à huile devienne la première plante oléagineuse au monde, avec une production annuelle d'huile de palme qui dépasse 70 Mt, à laquelle s'ajoute environ 8,5 Mt d'huile de palmiste (Figure 12) (USDA, 2022). En 2050, les besoins en huile de palme devraient se situer entre 120 et 156 Mt (Corley, 2009).

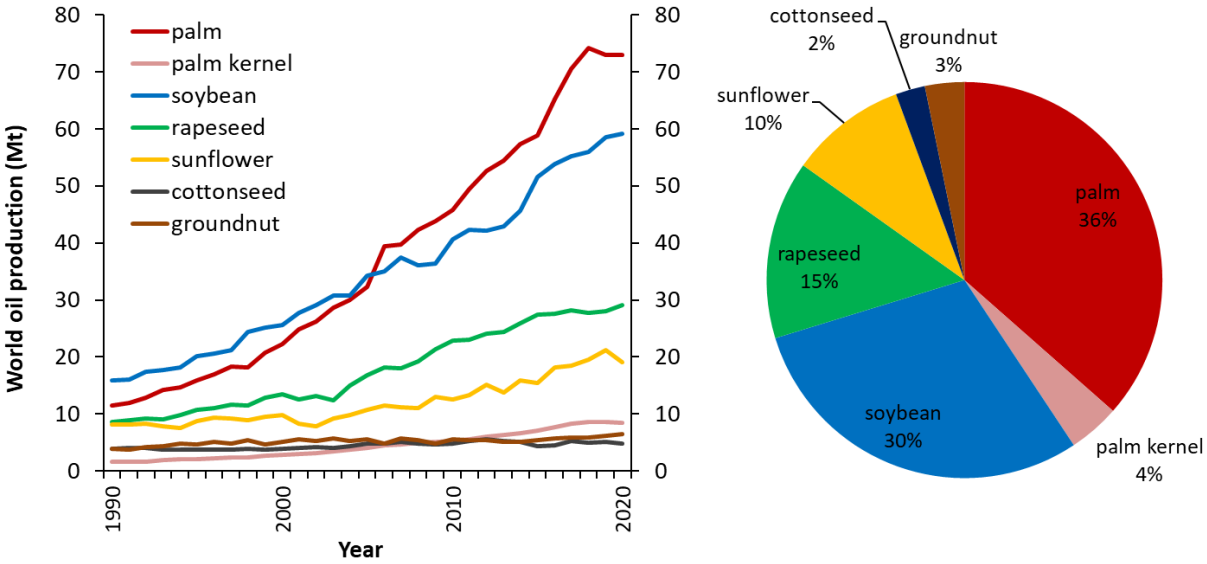


Figure 12 Evolution de la production des principales plantes oléagineuses depuis 1990 et part respective dans la production globale de 2021 (USDA, 2022)

La culture du palmier à huile s'étend sur plus de 20 Mha, répartis sur toute la zone de climat tropical. Les conditions de culture optimales sont une pluviométrie d'environ 1 800 mm par an, bien répartie le long de l'année, des températures minimales >18°C et un ensoleillement >1 800 heures par an. Ces exigences pédo-climatiques amènent une cohabitation forcée avec des zones de très forte biodiversité : Bornéo, Bassin du Congo, Amazonie. Les plantations de palmiers ont joué un rôle dans la déforestation de grandes étendues de forêts primaires ou secondaires, avec des conséquences très négatives (disparition d'habitats naturels, perte de biodiversité). Un des enjeux majeurs de la filière

palmier à huile aujourd'hui est donc l'évolution vers une production durable, avec une intensification sans polluer sur les surfaces existantes, afin de limiter le besoin en surfaces des nouvelles plantations et l'impact écologique de la culture. Des initiatives dans ce sens ont été prises. Ainsi, le RSPO (*roundtable on sustainable palm oil*, https://rspo.org/), créé en 2004, rassemble tous les acteurs de la filière au niveau mondial, depuis les producteurs jusqu'aux investisseurs financiers, et permet de définir et de promouvoir les règles d'une production durable d'huile de palme (Rival et Levang, 2013, p. 74). Il compte actuellement plus de 5 000 membres, dont le Cirad et PalmElit, et 3.34 millions d'hectares ont reçu une certification RSPO.

Les principaux producteurs sont l'Indonésie et la Malaisie, qui réalisent plus de 80% de la production mondiale (Figure 13). Le palmier à huile est un fort enjeu de développement pour de nombreux pays du Sud. Quand il est correctement planifié par les gouvernements et mis en œuvre par les planteurs, la culture du palmier à huile se traduit par un fort développement économique des régions concernées et par une importante réduction de la pauvreté rurale. Son exploitation repose sur des systèmes de culture très diversifiés allant de l'exploitation familiale de quelques hectares au périmètre agroindustriel de plusieurs dizaines (voire centaines) de milliers d'hectares. Plus de la moitié de l'huile de palme produite aujourd'hui provient de petites exploitations, au nombre d'environ trois millions. Les principaux consommateurs de l'huile de palme sont des pays émergents, l'Indonésie, l'Inde et la Chine étant les trois premiers.

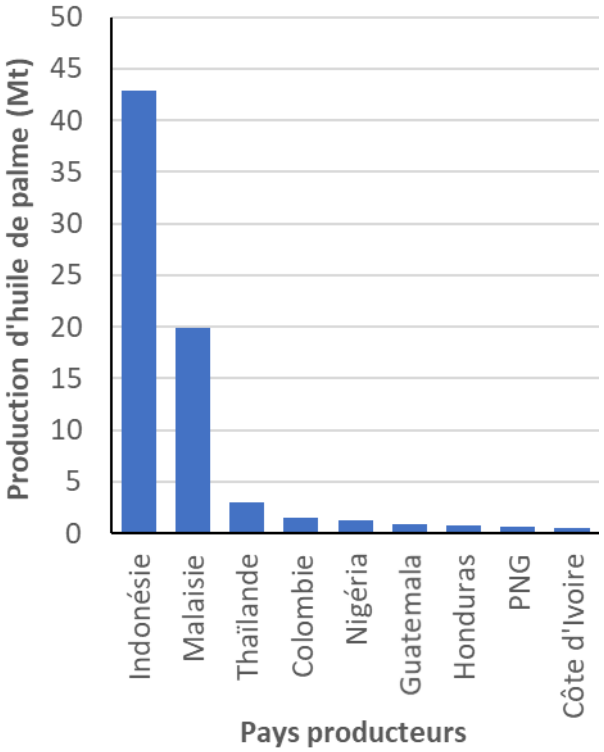


Figure 13 Principaux pays producteurs d'huile de palme (2019) (FAOSTAT, 2022)

e. Populations d'amélioration

Bien que l'utilisation du palmier à huile par les populations d'Afrique subsaharienne soit ancestrale, cette espèce n'a pas subi une domestication marquée et il n'existe pas de types distincts « sauvage » et « cultivé ».

Le palmier à huile a été introduit en Asie du Sud-est en 1848, avec quatre plantules dura plantées dans le jardin botanique de Bogor (Java, Indonésie) à des fins ornementales (Corley et Tinker, 2016). Leur origine exacte reste inconnue mais la population asiatique qui en a découlé (Deli) a des caractéristiques génétiques et phénotypiques proches de celles des populations d'Afrique centrale.

A partir des années 1920 une sélection massale a été appliquée au sein des différentes populations (Cochard, 2008 ; Corley et Tinker, 2016). En Afrique, elle a été réalisée principalement par les centres de recherches coloniaux (INEAC dans l'actuelle République Démocratique du Congo [RDC], IRHO en Côte d'Ivoire et au Bénin, WAIFOR au Nigéria). En Asie elle a été faite dans les grandes sociétés

de plantations d'Indonésie et de Malaisie. Ce processus s'est poursuivi jusque dans les années 1940 et a donné naissance aux populations d'amélioration modernes décrites ci-dessous, parfois nommées origines géographiques, origines génétiques ou, en anglais, BPRO pour *breeding populations of restricted origins*.

Les principales populations créées en Afrique sont La Mé (Côte d'Ivoire) et Yangambi (RDC). On trouve aussi les populations Yocoboué (Côte d'Ivoire), Sibiti (République du Congo), Ekona (Cameroun), WAIFOR (Nigéria) et Pobè (Bénin) mais elles sont beaucoup moins utilisées. La population La Mé trouve son origine dans les prospections faites dans la région de Bingerville dans les années 1920. Ceci a abouti à la sélection de 19 individus choisis car leurs fruits possédaient des proportions équilibrées entre le mésocarpe (60%), l'amande (20%) et la coque (20%). On note que ceci diffère notablement de l'idéotype moderne (Cochard, 2008). La population Yangambi est issue de plantations faites dans les années 1920 à partir de 10 à 20 tenera en pollinisation libre, incluant Djongo (« le meilleur ») du jardin botanique d'Eala et des tenera de Yawenda, Ngazi et Isangi. Une sélection a été faite sur la base du rendement en régimes puis de la qualité des régimes. Les objectifs étaient essentiellement une production de régimes élevée, un fort pourcentage de pulpe dans les fruits, de gros fruits et une amande relativement importante. Compte tenu des exigences élevées des sélectionneurs et des performances incomparables de Djongo, la population Yangambi d'origine serait issue à plus de 70% de Djongo. La population Sibiti est fortement apparentée à la population Yangambi, dont elle dérive (Cochard, 2008 ; Corley et Tinker, 2016 ; Demol et al., 2002).

En Asie, les quatre plantules de 1848 ont donné naissance à la population Deli, dans laquelle on distingue aujourd'hui plusieurs sous-populations, principalement Marihat Baris en Indonésie, SOCFIN en Indonésie et Malaisie, Serdang Avenue, Ulu Remis (ou Guthrie), Johor Labis et Elmina (dont les Dumpy) en Malaisie. Les premières activités connues de sélection de la population Deli pour le rendement en huile à partir d'observations rigoureuses datent des années 1910-1930, selon les sociétés de plantation (Cochard, 2008 ; Corley et Tinker, 2016). Les détails concernant cette période sont incertains (caractères sélectionnés, intensité de sélection, etc.).

Par ailleurs, des échanges de matériel ont abouti à la formation de l'origine Deli Dabou (Côte d'ivoire) à partir de graines de Deli SOCFIN et à la population AVROS (Indonésie, Malaisie) à partir de graines de Djongo.

Les populations de palmier à huile peuvent se répartir en deux groupes A et B selon les caractéristiques de production de leurs régimes (Gascon et de Berchoux, 1964). Le groupe A produit des régimes plus gros que le groupe B mais le groupe B produit un plus grand nombre de régimes. Le groupe A est composé des populations Deli et Angola, le groupe B des autres populations africaines. On peut à nouveau faire des distinctions entre populations du groupe B sur la base du phénotype, avec La Mé caractérisé par un nombre très faible de régimes et Yangambi par des régimes relativement gros. Les données moléculaires ont ensuite permis de préciser cette structure. Cochard (2008) et Cochard et al. (2009) ont étudié la diversité génétique sur un ensemble de 318 individus représentant huit pays, avec du matériel issu de prospections, de jardins botaniques et de programmes d'amélioration. Avec 14 marqueurs microsatellites (SSR), ils ont mis en évidence une structure très marquée (F_{ST} = 0.243), avec trois groupes de populations bien distincts :

- origines de Côte d'Ivoire,
- origines d'Afrique Centrale, du Nigéria et du Bénin,
- origine Deli.

f. Caractères cibles pour la sélection

L'amélioration génétique du palmier à huile vise plusieurs caractères. Les plus importants sont le rendement potentiel en huile de palme et la résistance aux maladies. D'autres caractères sont aussi considérés, notamment le développement végétatif (croissance en hauteur et encombrement), la résistance à la sécheresse et la qualité de l'huile (acidité de l'huile, profil d'acides gras, etc.). Tous ces caractères sont quantitatifs, à l'exception de l'acidité de l'huile (Domonhédo et al., 2018a).

Le rendement annuel en huile de palme d'un palmier est le produit du poids de sa production de régimes (PR) et du pourcentage d'huile dans ses régimes (%HR) (aussi nommé qualité des régimes, ou taux d'extraction). Le poids total de régimes est lui-même le produit du nombre de régimes (NR) et du poids moyen des régimes (PM). Il existe une corrélation négative forte entre NR et PM (Gascon et al., 1966). Le pourcentage d'huile dans les régimes est lui aussi le produit de caractères plus simples, qui sont le pourcentage de fruits dans le régime (%FR), le pourcentage de pulpe dans les fruits (%PF) et le pourcentage d'huile dans la pulpe fraîche des fruits (%HP). Cette décomposition correspond aux caractères sur lesquels portent actuellement la sélection pour l'amélioration du rendement.

g. Schéma d'amélioration

Des illustrations sur les activités d'amélioration génétique et de production de semences chez le palmier à huile sont fournies en Annexe 5.

Les inflorescences du palmier à huile, de grande taille, peuvent être isolées sur un individu et ensachées, ce qui permet de récolter du pollen et de réaliser des pollinisations contrôlées, selon des techniques développées dans les années 1940 (Annexe 5 N-P). Dans les conditions naturelles, le pollen est viable quelques jours, mais en le stockant sous vide au congélateur, il peut se conserver pendant plusieurs années. Ces caractéristiques du palmier à huile ont amené les programmes de sélection et de production de semences à s'appuyer exclusivement sur des croisements contrôlés.

A partir des années 1950, suite à la mise en évidence du déterminisme génétique du type de fruit (dura, tenera, pisifera), les palmiers à huile tenera ont remplacé les dura dans les plantations commerciales, amenant une augmentation de 30% du rendement en huile (Corley et Lee, 1992).

En 1957, la supériorité des hybrides A × B pour la production d'huile de palme a été mise en évidence (Figure 14). Elle résulte de la complémentarité entre les deux groupes pour les composantes de la production de régimes, qui permet aux croisements hybrides d'avoir une production annuelle de régimes dépassant de plus de 25% celle des populations parentales (Gascon et de Berchoux, 1964). Ceci a amené à l'adoption d'un schéma de sélection récurrente réciproque (SRR) (Gallais, 1990, p. 333-343 ; Gallais, 2009, p. 235 ; Gascon et de Berchoux, 1964 ; Meunier et Gascon, 1972), inspiré des travaux conduits chez le maïs (Comstock et al., 1949). Il utilise en général la population Deli pour le groupe A et La Mé ou AVROS pour le groupe B. Il permet, en utilisant des pisifera africains pour féconder les dura Deli, de produire des croisements commerciaux de type tenera, présentant de la vigueur hybride sur la production de régimes et un pourcentage élevé de pulpe dans les fruits.

Le détail du schéma est présenté Figure 15 (gauche). La population de départ de candidats à la sélection est composée d'individus appartenant, au sein des deux groupes, à des familles de pleins-frères. L'héritabilité au sens strict (h^2) des composantes du rendement en huile de palme se situe à des niveaux faibles à intermédiaires selon le caractère et la population (Corley et Tinker, 2016, p. 174, 180 ; Cros, 2014 ; Meunier et al., 1970). Les pourcentages de pulpe dans les fruits (%PF) et d'huile dans la pulpe (%HP), compte tenu de leurs valeurs de h^2 et de la relative facilité avec laquelle ils peuvent être évalués, sont utilisés pour appliquer une première étape de sélection dans laquelle les individus les moins performants sur ces caractères sont éliminés. Ceci ne s'applique par contre qu'aux dura et aux

tenera, les pisifera ne produisant pas de régimes. Cette première étape de sélection est rendue nécessaire par la lourdeur des évaluations en descendance hybride, qui empêche de tester tous les candidats à la sélection. Seuls les meilleurs pour %PF et %HP sont donc testés en croisement avec l'autre groupe, en général selon un plan de croisements de type NCM2 (*North Carolina model*) très incomplet, c-à-d avec en général deux à quatre croisements par parent. Les croisements sont observés dans des essais selon des dispositifs expérimentaux généralement de type blocs de Fisher ou lattice équilibré. Ces essais représentent des investissements lourds : actuellement chaque individu est croisé avec 2 à 4 partenaires, chaque croisement est représenté au champ par 45 à 72 individus et les observations sont réalisées de la 3ème à la 10ème année. A l'issue des essais, on obtient pour chaque parent une aptitude générale à la combinaison (AGC) avec de bonnes précisions, atteignant environ 0.90 pour toutes les composantes du rendement (Cros, 2014). Sur la base des AGC, la sélection finale est effectuée sur tous les caractères. Les aptitudes spécifiques à la combinaison (ASC) sont peu prises en compte, car elles sont beaucoup moins bien estimées que les AGC (précision autour de 0.3) et que le ratio entre la variance des ASC et la variance génétique totale entre croisements est faible (<15%) (Cros, 2014). Ce dernier point est probablement la conséquence de la SRR, qui fait diverger les fréquences alléliques entre groupes hétérotiques et réduit le ratio entre variance des ASC et variance des AGC au fil des cycles (Reif et al., 2007 ; Technow et al., 2014). Les individus sélectionnés produiront, par croisements au sein de chaque groupe et autofécondations, la génération suivante utilisée pour démarrer un nouveau cycle de SRR et pour produire du matériel commercial.

L'adoption dans les années 1950 de ce schéma de sélection a permis un progrès génétique important, estimé entre 1% et 1,5% par an (Durand-Gasselin et al., 2010 ; Rival et Levang, 2013). Cependant, l'obligation de conduire des tests sur descendance long et couteux a pour conséquence un intervalle de génération important (environ 20 ans) et une intensité de sélection faible (<200 individus testés en croisement par groupe).

Ce schéma permet aussi une sélection clonale au sein des croisements hybrides (Corley et Tinker, 2016, pp. 216-220).. Celle-ci est justifiée par la variabilité génétique générée par l'hétérozygotie existant dans les populations parentales, et qui atteint en moyenne 7% chez les Deli et 10% chez les La Mé du programme d'amélioration de PalmElit (Seyum et al., 2022b). Dans la perspective de sorties variétales clonales, les tenera présentant les meilleurs phénotypes sont choisis au sein des meilleurs croisements hybrides disponibles avant d'être évalués dans des essais clonaux. Les clones ont le potentiel d'augmenter encore le rendement du palmier à huile de 20 à 30% par rapport aux croisements sexuels (Corley et Law, 1997), et des augmentations de rendement de 13% (Nouy et al., 2006) et 18% (Soh et al. 2003a) ont été observées empiriquement. L'intérêt du clonage se réduit cependant au fil des cycles avec la baisse de l'hétérozygotie des populations parentales, mais la compréhension du mécanisme moléculaire à l'origine de l'anomalie associée à la culture *in vitro* devrait relancer l'intérêt de la sélection clonale.

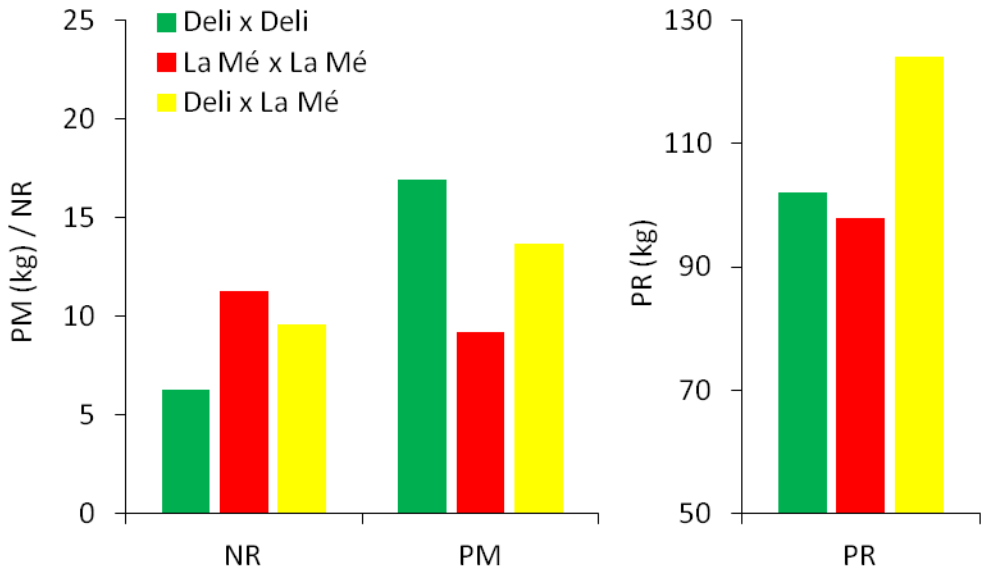
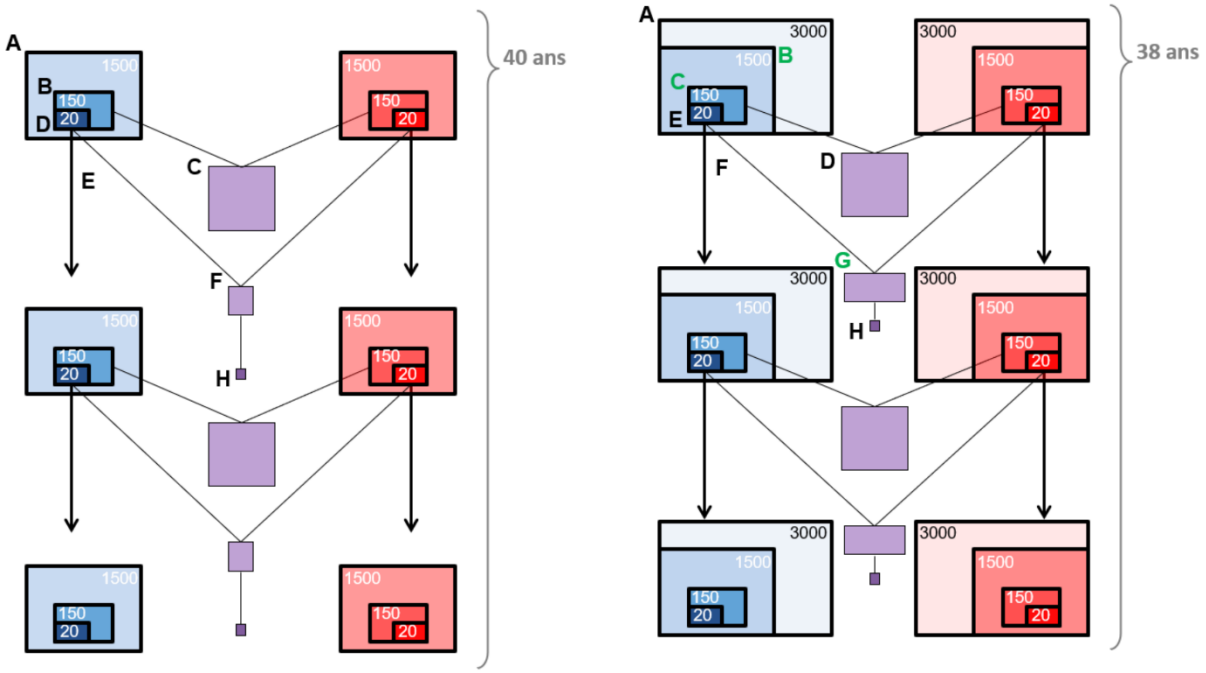


Figure 14 Production totale de régimes (PR) et ses composantes (nombre de régimes NR et poids des régimes PM) à l'âge adulte chez les dura des croisements intra- et inter-populations observés dans « l'Expérience Internationale », d'après les résultats donnés par Gascon et al. (1966)



Etapes de Sélection Récurrente Réciproque :

A champs semencier (candidats à la sélection phénotypique)
B sélection des individus à tester en croisements sur valeur propre des caractères à forte h^2
C tests en croisements hybrides
D sélection dans les individus testés en croisement (AGC/ASC)
E autofécondations et croisements des individus sélectionnés
F sélection phénotypique d'ortets dans des essais avec les meilleurs croisements
H essais clonaux et sélection parmi les clones testés

Etapes de Sélection Génomique (SG) Récurrente Réciproque :

A pépinière de candidats à la SG
B SG des individus qui constitueront le champ semencier, en pépinière, sur tous les caractères, à faible intensité, avec modèle calibré sur cycles précédents
C SG des individus à tester en croisements, sur tous les caractères, à forte intensité
D tests en croisements hybrides
E sélection dans les individus testés en croisement (AGC/ASC)
F autofécondations et croisements des individus sélectionnés
G SG d'ortets dans une pépinière des meilleurs croisements hybrides avec modèle calibré sur cycles précédents et en cours
H essais clonaux et sélection parmi les clones testés

Le nombre d'individus à chaque étape est indicatif. En vert les étapes ou la SG est appliquée (dans les deux populations parentales).

Figure 15 Schéma d'amélioration phénotypique par sélection récurrente réciproque du palmier à huile (gauche) et alternative génomique (droite)

2.2.5. L'hévéa et son amélioration génétique

a. La plante

L'hévéa (*Hevea brasiliensis* Müll. Arg.) est un arbre décidu de la famille des Euphorbiacées, appartenant aux dicotylédones, et est originaire de la forêt Amazonienne (Clément-Demange et al., 2007).

Il peut atteindre 30 mètres en plantation (Figure 16). Il a une croissance rythmique, avec une alternance entre une phase d'élongation rapide d'un entre-nœud et une période de repos pendant laquelle les feuilles autour du bourgeon terminal se développent. Il porte des feuilles trifoliées caractéristiques. Des illustrations sur la plante, sa filière et son amélioration génétique sont fournies en Annexe 6.

Figure 16 Hévéa en plantation

L'hévéa commence à produire des inflorescences vers quatre ou cinq ans. Il est monoïque, avec des inflorescences incluant des fleurs séparées mâles et femelles. L'hévéa est préférentiellement allogame en raison d'une certaine protandrie (maturité plus précoce des fleurs mâles) et d'une forte auto-incompatibilité. Les inflorescences apparaissent à la fin du processus de défoliation-refoliation qui se produit durant la saison sèche. La pollinisation est entomophile, et implique en particulier des mouches et des moucherons. La multiplication végétative est couramment utilisée chez l'hévéa, une méthode efficace de greffe de bourgeons ayant été développée dans les années 1910 (Annexe 6 I et J).

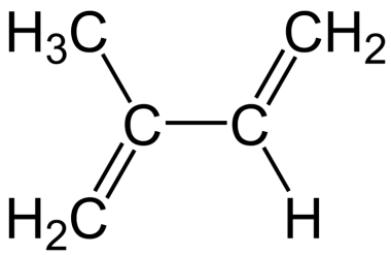


Figure 17 Molécule d'isoprène (C_5H_8)

L'hévéa est diploïde et possède 18 paires de chromosomes (2n=36). Son génome couvre une distance génétique d'environ 2 250 cM, et une distance physique de 2,1 Gb (Munyengwa et al., 2021). Une seule séquence du génome nucléaire assemblée en pseudo-chromosome est disponible (Liu et al., 2020).

Les principaux stress biotiques auxquels est soumis l'hévéa sont des maladies foliaires fongiques, qui peuvent avoir une incidence économique très forte. Il s'agit en particulier de la maladie Sud-américaine des feuilles (*South American Leaf Blight*, SALB), causée par *Microcyclus ulei*, et les maladies dues à *Corynespora cassiicola* et à *Colletotrichum gloeosporioides*, en Afrique et en Asie (Garcia, 2017 ; Pujade-Renaud, 2015).

L'hévéa peut aussi être soumis à un désordre physiologique, le syndrome de l'encoche sèche (TPD, *tapping panel dryness*). Il correspond à l'assèchement spontané de l'encoche réalisée lors de la

saignée destinée à collecter le latex (voir 2.2.5.d). Le TPD cause une réduction du rendement, voire un arrêt de la production.

b. Le caoutchouc naturel

L'hévéa fait partie des plantes produisant du latex. Selon une conception évolutive, le métabolisme de biosynthèse du latex serait un mécanisme naturel de défense, notamment vis-à-vis des blessures de l'écorce. Le latex est synthétisé et stocké dans le cytoplasme de cellules spécialisées, les laticifères, situées dans le liber (Annexe 6 E). Après quelques mois de croissance des arbres, les cellules laticifères fusionnent pour donner un réseau para-circulatoire de « vaisseaux laticifères », ce qui permettra l'écoulement et la récolte du latex lors de la saignée.

Le latex est une suspension colloïdale blanche contenant 90% de particules de caoutchouc. Le caoutchouc naturel, ou cis-1,4-polyisoprène, est un polymère d'isoprène (2-méthyl-1,3-butadiène, Figure 17), contenant 150 à 2 000 000 de molécules d'isoprène. Il y a deux classes de particules de caoutchouc dans le latex d'hévéa : les grosses particules de caoutchouc, représentant plus de 90% du caoutchouc en volume dans le latex, et les petites particules de caoutchouc.

Le caoutchouc naturel, élastique, devient cassant à froid et collant à chaud. En 1839, la découverte par Charles Goodyear de la vulcanisation, un procédé consistant à chauffer le caoutchouc en présence de soufre (Goodyear, 1853), a joué un grand rôle dans l'histoire de la filière. Ce traitement permet de durcir le caoutchouc en formant des liaisons transversales entre les sections de la chaîne polymère, ce qui accroit la rigidité et la durabilité. Ceci assure la stabilisation de ses propriétés sur une large plage de température et son utilisation industrielle.

c. Utilisations

L'hévéa est la seule source économique viable de caoutchouc naturel, de par son rendement et les excellentes propriétés physiques de son caoutchouc. Environ 70% du caoutchouc naturel est utilisé dans l'industrie des pneumatiques (voiture, avion, vélos), en raison de son élasticité et de sa solidité. Le caoutchouc naturel est concurrencé depuis 1945 par les caoutchoucs synthétiques (principalement styrène-butadiène), produits à partir du pétrole et du gaz naturel. Il conserve cependant en 2022 une part d'utilisation de 47 % sur le marché des élastomères en raison de ses propriétés particulières : cristallisation sous tension permettant un durcissement à l'échauffement et une forte résistance à la déchirure, et pouvoir collant élevé permettant une bonne liaison entre les différentes couches de produits lors de la fabrication des pneus. Ces propriétés le rendent indispensable en particulier pour les pneus d'avions, de camions et d'engins de génie civil (Vaysse et al., 2012).

Au moment de l'abattage des parcelles, les arbres fournissent aussi du bois, un produit secondaire qui représente en Asie environ 15 % de la valeur productive de ces parcelles.

d. Filière et production

Après une phase de croissance dite « immature » conduite jusqu'à la fermeture de la canopée, les arbres sont mis en saignée (« ouverture ») lorsque la circonférence des troncs atteint 50 cm à une hauteur de 125 cm, soit entre cinq et sept ans. La collecte du latex se fait par la saignée (voir Annexe 6 F). Celle-ci consiste à creuser une encoche dans le tronc, de façon à atteindre le liber mais sans blesser le cambium. La saignée sectionne les laticifères, dont le contenu cytoplasmique est expulsé, formant un écoulement de latex. Ce latex est ensuite coagulé et séché dans des usines proches des

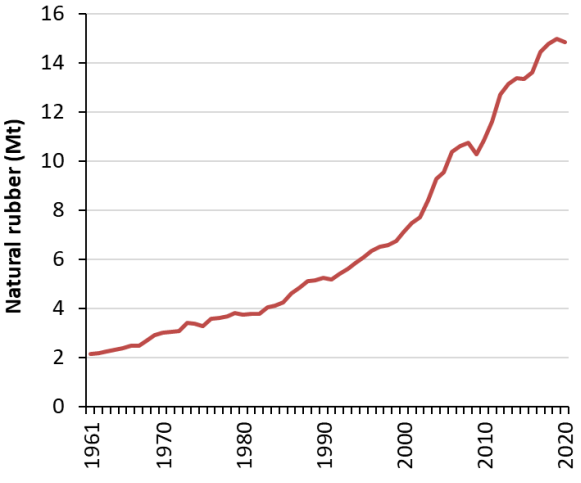
plantations. Une parcelle est saignée tous les 2 à 7 jours pendant 25 à 30 ans, car au-delà la production diminue et la saignée n'est plus rentable. La densité de plantation est normalement de 500 arbres par hectare.

La découverte de la stimulation par l'éthylène (Abraham, 1968) a ouvert de nouvelles perspectives pour l'exploitation des hévéas, permettant d'augmenter la production ou de réduire la fréquence de saignée, et donc d'augmenter la productivité du travail. La biosynthèse du caoutchouc naturel est influencée par diverses hormones végétales. L'éthylène a été identifié comme stimulant la production de latex, et est appliqué sous forme d'éthéphon (un libérateur d'éthylène). Le traitement de l'écorce par l'éthéphon augmente le rendement en latex de 1,5 à 2 fois.

Le développement du caoutchouc naturel est intrinsèquement lié au développement de l'automobile et des engins roulants sur pneus, avec une augmentation de la production mondiale qui a été accélérée après 1990 par la mondialisation et le développement de la région Asie-Pacifique, notamment de la Chine. La production mondiale annuelle de caoutchouc naturel dépasse 14 Mt (Figure 18). Plus de 90% de la production se fait en Asie. Les principaux producteurs sont la Thaïlande et l'Indonésie (>50% de la production mondiale) (Figure 19). Bien que l'hévéa soit originaire d'Amazonie, l'Amérique du Sud représente une faible part dans la production à cause du SALB (Garcia, 2017).

L'hévéa est cultivé dans toute la zone de climat tropical, avec une surface mondiale d'environ 15 Mha. La surface cultivée appartient en grande partie à des petits producteurs villageois (85%). Les zones les plus favorables sont proches de l'équateur, et les zones proches des tropiques et celles affectées par le SALB correspondent à des régions de culture marginales.

Les prévisions indiquent que la demande de caoutchouc naturel dépassera 19 Mt en 2025 (Warren-Thomas et al., 2015), alors que les plantations d'hévéas sont déjà responsables de déforestation et constituent des menaces pour la biodiversité, notamment en Asie du Sud-Est (Ahrends et al., 2015 ; Warren-Thomas et al., 2015). Une démarche de durabilité vise donc à freiner l'extension des surfaces sous hévéa, tout en cherchant à répondre à la demande par une augmentation de la productivité des surfaces existantes avec des techniques préservant la fertilité des sols. La productivité du travail est également importante pour l'amélioration des revenus des petits planteurs et des travailleurs salariés. Ces enjeux concernent fortement l'amélioration génétique pour l'obtention de clones d'hévéa performants et adaptés à l'évolution des contextes économique et climatique.



http://www.fao.org/faostat/en/#data/QC - 2022, March 25

Figure 18 Evolution de la production mondiale de caoutchouc naturel depuis 1961

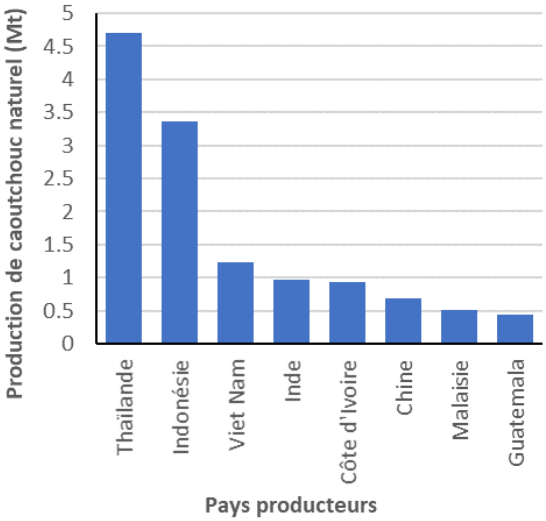


Figure 19 Principaux pays producteurs de caoutchouc naturel (2020) (FAOSTAT, 2022)

e. Populations d'amélioration

Au 19ème siècle, le caoutchouc naturel est récolté dans les forêts amazoniennes par les *seringueiros*. L'administration coloniale britannique cherche à mettre en culture l'hévéa en Asie. En 1876, 70 000 graines sont collectées par Henry Wickham dans la forêt tropicale brésilienne et envoyées en Angleterre, à Kew Gardens. Ces graines ont donné 2 700 plantules qui ont été envoyées en Asie du Sud-Est (Sri-Lanka et Singapour), où elles ont été à la base de la population domestiquée dite Wickham. La plupart des cultivars améliorés provenant des centres de sélection asiatiques ou africains appartiennent à cette population.

Les ressources génétiques de l'hévéa incluent, en dehors des Wickham, des accessions sauvages d'Amazonie. Elles sont organisées selon trois groupes, Acre, Rondonia et Mato Grosso, correspondant aux bassins hydrographiques des principaux affluents de l'Amazone (Le Guen et al., 2009).

f. Caractères cibles pour la sélection

Les critères classiques de sélection sont : la production à court terme (≤10 ans), moyen et long terme ; la résistance aux facteurs de réduction du peuplement saigné, à savoir la résistance au TPD (2.2.5.a), et la résistance aux dommages dus au vent ; la résistance aux maladies fongiques de feuilles (SALB, *Corynespora* et *Colletotrichum*) ; la vitesse de croissance immature (qui définit la précocité de la production) ; l'aptitude à la croissance en cours de saignée ; et le taux de saccharose du latex, un précurseur indispensable de la biosynthèse du caoutchouc. Tous ces caractères sont quantitatifs.

La sélection pour la résistance aux maladies fongiques de feuilles est compliquée par la spécificité des méthodes de phénotypage de la résistance, qui impose généralement des essais particuliers, et par la difficulté d'évaluer précocement des clones pour la résistance aux différents pathotypes, certains d'entre eux pouvant émerger sur certains clones de façon tardive dans le processus de sélection.

Des caractères moins importants sont aussi visés, comme la qualité technologique du caoutchouc naturel, la production de bois et l'aptitude au greffage.

La sélection est réalisée sur un indice multi-caractère, en équilibrant le poids des différents caractères en fonction de leur importance agronomique. Pour un clone *i*, il vaut : $I_i = \sum_{t=1}^{n} \alpha_t g_{t_i}$, avec n le nombre de caractères, α_t le poids associé au caractère t et g_{t_i} la valeur génétique estimée du clone *i* pour le caractère t.

g. Schéma d'amélioration

Il est possible de réaliser des croisements contrôlés entre hévéas. Cela passe par la collecte d'une fleur mâle chez un premier clone, l'extraction de la colonne staminale et son insertion entre les pétales fermés d'une fleur femelle d'un second clone, qui sera scellée avec du coton. Cependant, produire de cette façon une quantité suffisante de graines d'un croisement particulier est un travail complexe. Par ailleurs, il existe de grandes variations entre clones en termes de floraison, de fertilité et de fructification, allant d'une quasi stérilité à une fertilité importante. Ceci représente une limitation forte à l'utilisation de croisements contrôlés dans l'amélioration de l'hévéa, et rend notamment difficile la réalisation de plans de croisements structurés pour l'évaluation des valeurs en croisements. Par ailleurs, de très importants efforts ont été consacrés à la mise au point d'une méthode efficace de multiplication végétative. Le bouturage n'a pas permis d'obtenir des systèmes racinaires profonds et

performants, et les essais de plantations de boutures ont conduit à des déracinements massifs sous l'effet du vent. La culture *in vitro* n'a pas non plus abouti à des résultats probants. La multiplication par greffe de bourgeons (1917) permet par contre, selon des techniques horticoles simples (Annexe 6 I et K), d'assurer la multiplication des individus élites à grande échelle. Les contraintes biologiques de l'hévéa et la mise au point de la greffe de bourgeons ont abouti à ce que les plantations clonales remplacent progressivement les plantations issues de semis. Aujourd'hui, tous les programmes d'amélioration de l'hévéa visent des sorties variétales clonales. Il est cependant important de souligner que l'amélioration génétique ne porte dans ce cas que sur la partie aérienne des arbres, le système racinaire restant très peu sélectionné car constitué de populations de demi-frères issues d'un clone jugé bon grainier et porteur d'aptitudes favorables pour la partie racinaire.

La sélection clonale est conduite par familles de plein-frères, avec des croisements contrôlés réalisés de façon ponctuelle, pour générer des familles biparentales qui servent de population de sélection, et avec un large recours à la multiplication végétative pour les évaluations et la diffusion des clones sélectionnés.

Ce schéma de sélection clonale comporte quatre étapes principales (Figure 20, gauche) :

- La réalisation d'un croisement biparental,
- La conduite d'un essai d'évaluation de jeunes plants (SET, *seedling evaluation trial*). Il est réalisé sur un grand nombre d'individus (~3 000) non bouturés (c-à-d avec une copie par génotype), plantés à forte densité. Il dure deux ou trois ans. En SET, seule la production de latex s'avère suffisamment héritable (au sens large) et dotée d'une bonne valeur prédictive du comportement des arbres greffés pour permettre une sélection efficace.
- La conduite d'un essai d'évaluation de clones à petite échelle (SSCT, *small scale clonal trial*). Il est réalisé sur un nombre restreint de clones (<200) représentés par 10 à 30 ramets par clone, plantés en général selon un dispositif expérimental avec blocs complètement randomisés. Il dure quatre à huit ans, avec une densité de plantation élevée ou normale. Le CCPE permet d'étudier les principaux caractères liés aux arbres individuels (croissance immature, production de latex, taux de saccharose dans les laticifères, abondance de branchement), et de réaliser une première évaluation de la tolérance aux maladies de feuilles.
- La conduite d'un essai d'évaluation de clones à grande échelle (LSCT, *large scale clonal trial*). Il est réalisé sur un très petit nombre de clones (10-20) représentés par plusieurs centaines de ramets, sur une longue durée (≥15 ans) et en multi-local, avec un dispositif expérimental en blocs complétement randomisés et une densité normale. Il permet une évaluation agronomique en conditions réelles sur l'ensemble des caractères d'intérêt (production, croissance, résistance aux stress biotiques et abiotiques, réponse à la stimulation par l'éthylène, etc.).

L'amélioration génétique de l'hévéa a permis une augmentation importante de la production. L'évaluation des accessions amazoniennes sauvages issues de prospections, avec les méthodes modernes de l'hévéaculture, a permis d'estimer la production de latex de ces populations à environ 300 kg/ha/an. Aujourd'hui, les meilleurs clones produisent plus de 2 500 kg de latex par hectare et par an (Priyadarshan, 2017, p. 110).

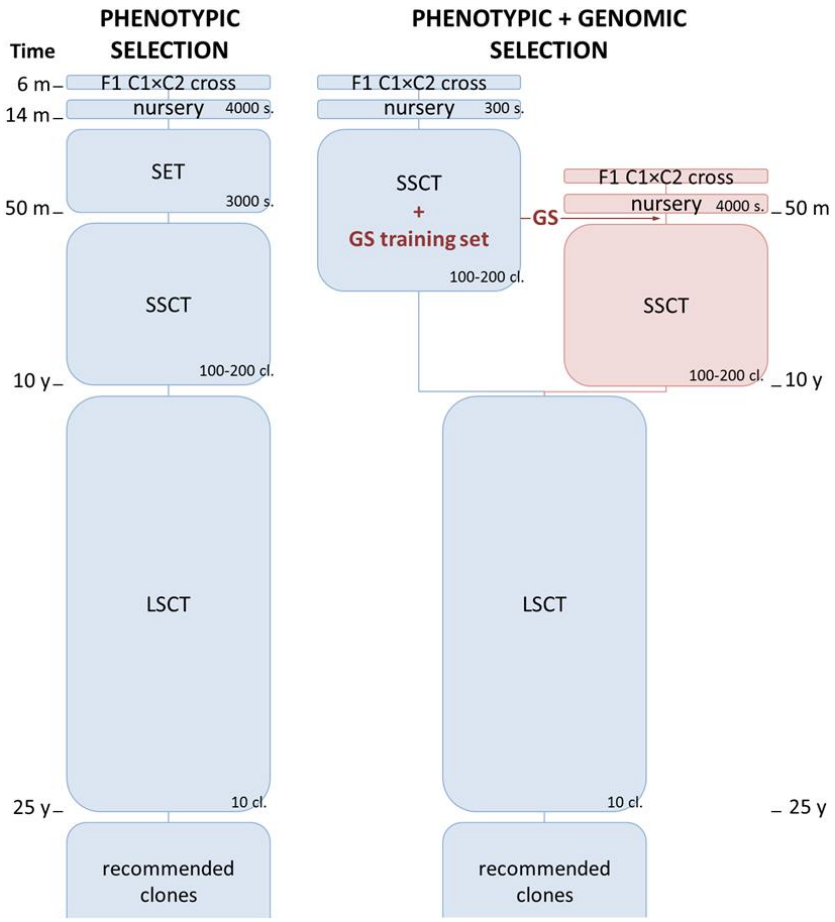


Figure 20 Schéma d'amélioration phénotypique de l'hévéa (gauche) et alternative génomique (droite) à partir d'un croisement biparental C1xC2 (Cros et al., 2019)
SET : seedling evaluation trial; SSCT, small scale clonal trial, LSCT; large scale clonal trial

2.3. Facteurs influençant la précision des prédictions génomiques

Un de mes deux principaux sujets d'étude est la compréhension des facteurs influençant la précision des prédictions génomiques dans le contexte des populations d'amélioration du palmier à huile et de l'hévéa.

La précision de la SG est affectée par de nombreux facteurs : la taille efficace de la population (*Ne*), le déséquilibre de liaison (DL) entre marqueurs et QTL, la densité et le type de marqueurs, la taille et la structure de la population de calibration, l'architecture génétique des traits, la parenté entre la population de calibration et la population de sélection, l'héritabilité des caractères, la méthode d'imputation des données moléculaires, etc. (Grattapaglia et Resende, 2011 ; Isik, 2014 ; Robertsen et al., 2019).

Je présente ci-dessous les principaux facteurs dont j'ai étudié l'effet. J'ai opté pour une présentation séquentielle pour des raisons de clarté, mais la plupart des facteurs sont interconnectés et leurs effets ne sont pas indépendants.

2.3.1. Approches statistiques de prédiction et architecture génétique des caractères

a. Méthodes statistiques de prédiction

La performance relative des différentes méthodes statistiques est supposée varier en fonction de l'architecture génétique du caractère considéré (Lebedev et al., 2020 ; Lorenz et al., 2011, p. 92). L'architecture génétique correspond aux caractéristiques génétiques qui déterminent la relation génotype-phénotype, et en particulier le nombre de gènes qui contrôlent le caractère, le nombre d'allèles par gène, la distribution des gènes le long du génome, la distribution des effets des gènes et le mode d'action des gènes (additif, dominant, épistatique) (Momen et al., 2018).

Ainsi, les méthodes dans lesquelles les effets des marqueurs sont échantillonnés dans des distributions où la variance est la même pour tous les marqueurs, comme le GBLUP (VanRaden, 2008 ; VanRaden, 2007), le RRBLUP (Meuwissen et al., 2001) et la *Bayesian ridge regression* (Pérez et al., 2010), devraient être plus appropriées pour les caractères qui suivent le modèle infinitésimal. Au contraire, les méthodes avec des variances spécifiques aux marqueurs, comme le LASSO Bayésien (de los Campos et al., 2009), BayesA, BayesB (Meuwissen et al., 2001), BayesCπ et Bayes Dπ (Habier et al., 2011), devraient être plus appropriées pour les caractères dont l'architecture génétique comprend des QTL majeurs. Par conséquent, de nombreuses études de SG ont comparé des méthodes statistiques de prédiction afin d'identifier la plus appropriée pour un caractère donné.

Sur le palmier à huile, j'ai mené ce type de comparaison, en considérant le GBLUP, le LASSO Bayésien, la *Bayesian random regression*, BayesCπ et Bayes Dπ (Cros et al., 2015b). Sur l'hévéa, les méthodes RRBLUP, LASSO Bayésien et RKHS ont été comparées lors des stages de M2 de Luther Nkoulou et Jean Oum (Cros et al., 2019). Chez les deux espèces, peu de variations ont été trouvées entre méthodes statistiques et n'étaient pas significatives, comme on peut le voir sur la Figure 21 avec l'exemple de la production de latex chez l'hévéa. Chez le palmier à huile, cet aspect a aussi été étudié par d'autres auteurs, qui ont conclu que certaines méthodes pouvaient être plus précises, en l'occurrence BayesB chez Ithnin et al. (2017) et le *support vector machine* (SVM), une méthode de *machine learning* (Long et al., 2011), chez Kwong et al. (2017b). Cependant, ces études mettaient en évidence des variations de faible magnitude. Par exemple, chez Kwong et al. (2017b), la précision de prédiction moyenne sur six caractères était de 0.33 pour le SVM, contre 0.29 à 0.31 pour les sept autres méthodes considérées, avec des changements de rangs entre méthodes en fonction des caractères, une population de petite taille (112 individus) et pas de tests statistiques pour évaluer si les différences entre méthodes étaient significatives. Dans ces conditions, on peut se questionner sur la validité des différences. Dans mes études, j'ai utilisé des tests statistiques avant de conclure quant à l'effet sur la précision de la SG des différents facteurs considérés. Dans les premiers articles, j'ai procédé par analyse de variance ou par tests *t* de Student apparié, avec au préalable une transformation des précisions en variable *Z* de Fisher, et en considérant des répétitions de validation (obtenues pour les analyses sur un seul site, par le processus de validation croisée, et, pour les analyses entre site, en découpant la population de validation) (Cros et al., 2019 ; Cros et al., 2018). Plus récemment (Munyengwa et al., 2021 ; Nyouma et al., 2020), j'ai utilisé le test *t* d'Hotelling-Williams (Steiger, 1980). Ce test compare deux coefficients de corrélation de Pearson dépendants, c-à-d ayant une variable en commun. Dans le cas de la précision de SG, cette variable en commun correspond aux valeurs génétiques de la population de validation (valeurs réelles, dans les études par simulations, ou, plus généralement, observées ; voir 2.2.3.b). Ce test est particulièrement intéressant pour les validations entre dispositifs car il ne nécessite pas de découper la population de validation en répétitions : ainsi, les comparaisons se font sur une valeur de précision plus fiable car calculée sur la population de

validation dans son ensemble, et elles ne dépendent plus d'éventuels effets aléatoires liés au découpage de la population de validation.

Les faibles variations observées sur les précisions de SG en fonction des approches statistiques suggèrent que les caractères que j'ai étudiés sont largement polygéniques et suivent le modèle infinitésimal (très grand nombre de gènes d'effets très faibles). Ces résultats confirment ceux obtenus dans des évaluations empiriques d'autres espèces, où les différentes approches tendaient à avoir des performances similaires. Ceci a abouti à ce que le GBLUP, qui est computationnellement plus simple à mettre en œuvre, devienne la méthode la plus largement utilisée en prédictions génomiques (Heslot et al., 2015 ; Montesinos-López et al., 2021).

b. Modélisation des effets non-additifs

De la même façon, les modèles tenant compte des effets génétiques non-additifs peuvent potentiellement augmenter la précision des prédictions génomiques pour les caractères dont le déterminisme génétique n'est pas purement additif.

Sur le palmier à huile, avec le GBLUP, nous avons comparé des modèles purement additifs et des modèles incluant aussi des effets non-additifs (Cros et al., 2017 ; Nyouma et al., 2020). Sur l'hévéa, le RRBLUP et le LASSO Bayésien ont été implémentés avec des modèles purement additifs et des modèles incluant aussi des effets de dominance ; et nous avons utilisé le RKHS, qui inclut implicitement les effets non-additifs (Cros et al., 2019).

Chez les deux espèces, peu de variations ont été trouvées entre modèles à effets additifs seuls et modèles incluant aussi des effets non-additifs, et elles n'étaient pas significatives (voir exemple Figure 21).

Les études conduites chez d'autres espèces suggèrent que le fait que les modèles intégrant des effets non additifs n'aient pas augmenté les précisions dans mes études sur le palmier et l'hévéa pourrait avoir deux causes (Cros et al., 2017 ; Cros et al., 2019 ; Nyouma et al., 2020) :

- une part insuffisante de ce type d'effet dans la variance génétique. Dans une étude par simulation sur l'eucalyptus, Denis et al. (2013) ont par exemple montré que les modèles avec effets de dominance amélioraient la précision de la SG si le ratio de variance de dominance sur variance additive était au moins de 1. Les résultats chez le palmier à huile sont donc cohérents avec une études antérieure (Cros, 2014), qui a mis en évidence dans le même dispositif expérimental un ratio entre la variance des ASC et la variance génétique entre croisements <15%,
- une population de calibration de taille insuffisante (au moins plusieurs centaines d'individus seraient nécessaires).

Pour le palmier à huile, cet aspect mériterait toutefois d'être réexaminé en utilisant la modélisation des effets de dominance développée par González-Diéguez et al. (2021). Dans les études Cros et al. (2017) et Nyouma et al. (2020), et comme dans de nombreux articles chez d'autres espèces, notamment chez le maïs (González-Diéguez et al., 2021), je me suis basé sur l'approche de Stuber et Cockerham (1966) pour modéliser les ASC. Dans cette approche, la matrice des apparentements génomiques entre croisements hybrides associée aux ASC est obtenue en faisant le produit de Kronecker des matrices d'apparentements génomiques additifs des deux groupes parentaux. González-Diéguez et al. (2021) ont montré que, dans un contexte de prédictions génomiques, cette approche ne permettait pas de capter tous les effets non additifs, et ils ont proposé une modélisation alternative décomposant la dominance, l'épistasie additive intra- et inter-populations parentales,

l'épistasie entre les effets additifs des deux populations parentales et les effets de dominance, et l'épistasie de dominance.

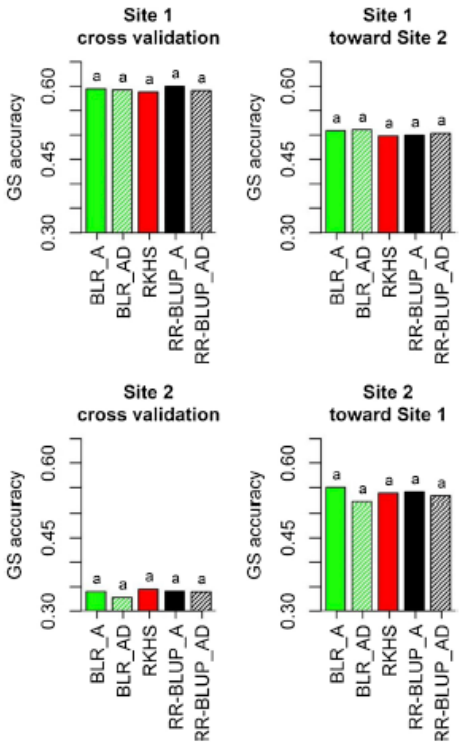


Fig. 2. GS accuracy for rubber production according to statistical method of GS prediction, and validation approach. Values are means over seven replicates for Site 1 cross validation and Site 2 to Site 1 independent validation, and five replicates for Site 2 cross validation and Site 1 to Site 2 independent validation. Values with the same letter within a given validation approach are not significantly different at P = 0.05. All the clones were used to train the GS model. All the SSRs were used.

Figure 21 Comparaison de méthodes statistiques de prédiction pour la production de latex dans une famille de plein-frères d'hévéa, avec des validations croisées et des validations entre sites (Cros et al., 2019)

c. Modèles pour l'évaluation d'hybrides

Sur le palmier à huile, les croisements évalués étant des hybrides entre deux sources génétiquement distinctes, des aspects particuliers peuvent être pris en compte en termes de modélisation (Figure 22).

Lorsque seuls les parents des croisements hybrides sont génotypés, on applique un modèle dit parental, dans lequel on décompose la valeur génétique additive des croisements en deux parties indépendantes héritées de chacun des deux groupes

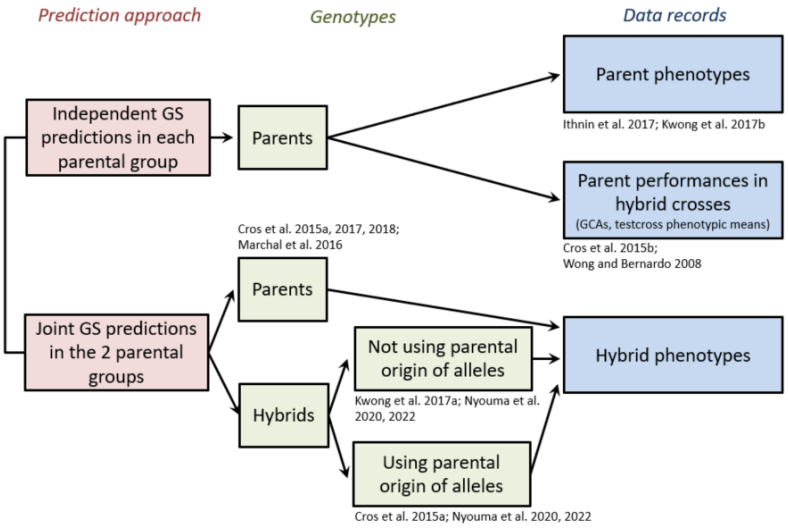


Figure 22 Récapitulatif des différentes approches de modélisation étudiées chez le palmier à huile pour les prédictions génomiques

parentaux, les aptitudes générales à la combinaison (AGC) (Lo et al., 1997 ; de Souza Jr, 1992 ; Stock et al., 2020 ; Stuber et Cockerham, 1966).

Une possibilité pour réaliser des prédictions génomiques est d'appliquer un modèle indépendamment par groupe parental, en utilisant les AGC comme phénotypes. Cette approche est avantageuse car son temps de calcul est très réduit, ce qui m'a amené à la choisir dans mon premier article (Cros et al., 2015b). Une approche similaire avait par ailleurs été choisie par Wong et Bernardo (2008), qui avaient utilisé comme phénotypes les moyennes phénotypiques des descendants en *testcross*, dans la première étude consacrée à la SG chez le palmier à huile.

D'autres auteurs ont mis en œuvre cette approche chez le palmier à huile dans une version utilisant les valeurs propres individuelles comme phénotypes (Ithnin et al., 2017 ; Kwong et al., 2017b). Cependant, les phénotypes parentaux peuvent ne pas refléter les performances dans les croisements hybrides en raison des différences de fréquence des gènes entre les populations parentales et des effets non additifs (Baumung et al., 1997 ; Stock et al., 2020 ; Vitezica et al., 2016 ; Wei et al., 1991), si bien qu'on peut questionner la pertinence de ce type d'approche.

Par la suite, j'ai préféré aller vers une modélisation plus complexe, c-à-d réalisant conjointement les deux étapes d'analyses précédemment séparées (estimation des AGC à partir des phénotypes hybrides, puis utilisation dans un modèle de prédiction génomique) (Cros et al., 2017 ; Cros et al., 2018 ; Marchal et al., 2016 ; Nyouma et al., 2022 ; Nyouma et al., 2020). Ceci permet de prédire les GEBV des deux groupes parentaux avec une seule analyse, et offre la possibilité de prédire les ASC des croisements, même s'il s'est avéré que cela n'impactait pas les précisions. Par ailleurs, il n'est plus nécessaire d'utiliser des variances résiduelles hétérogènes comme dans l'approche précédente, les variations en termes de qualité d'évaluation des différents parents étant prises en compte par le modèle. Il n'est plus non plus nécessaire de dérégresser les données phénotypiques, ce qui doit être fait lorsque celles-ci ont été obtenues par la méthodologie du BLUP (Cros et al., 2015b, p. 400). Enfin, les individus des groupes A et B n'étant pas homozygotes, cette approche donne aussi la possibilité d'utiliser les données moléculaires des individus hybrides, afin de tirer profit de la ségrégation existant au sein des croisements. Ceci permet d'augmenter la taille de la population de calibration (voir 2.3.4.b), et donne aussi la possibilité de deux modélisations alternatives au modèle parental (Ibánez-Escriche et al., 2009 ; Stock et al., 2020) :

(i) PSAM, pour *population* (ou *breed*, dans le context animal) *specific effects of single nucleotide polymorphism alleles model*, dans lequel les effets aux marqueurs sont spécifiques de la population parentale

(ii) ASGM, pour *across-population SNP genotype model*, qui estime un seul effet par marqueur.

Chez le palmier à huile, nous avons mis en évidence que le modèle ASGM était plus performant, car il donnait une précision de prédiction légèrement plus élevée en moyenne sur l'ensemble des caractères, était le meilleur modèle sur un plus grand nombre de caractères et avait des performances moins affectées par la population et le jeu de SNP (Nyouma et al., 2022 ; Nyouma et al., 2020). La Figure 23 montre un exemple des résultats obtenus. Elle présente la précision moyenne sur neuf composantes du rendement en huile de palme avec différents modèles. On voit que, quel que soit le modèle, l'approche ASGM donne des précisions supérieures ou égales à PSAM. Technow et al. (2012) ont signalé que PSAM prédisait mieux les valeurs génétiques en cas de faible persistance des phases entre populations parentales, ce qui implique que la performance relative de PSAM et ASGM est affectée par la densité de marqueurs et par l'histoire des populations parentales, et notamment le nombre de générations depuis leur divergence. Chez les animaux, PSAM est apparu plus précis que ASGM pour une faible densité de marqueurs (400), une grand population de calibration (4 000) et un

faible apparentement entre races (origine commune remontant à au moins 550 générations) chez Ibánez-Escriche et al. (2009) ; et pour des parents hybrides génétiquement distants, c'est-à-dire ayant divergé il y a 300-400 générations, et une grande population d'entraînement avec 2 000 à 8 000 individus chez Esfandyari et al. (2015).

Suite à ces résultats concernant les performances relatives de PSAM et ASGM chez le palmier à huile, obtenus durant le doctorat d'Achille Nyouma, j'ai inclus une étude de génétique des populations dans le travail de doctorat d'Essubalew Seyum. Il a étudié, sur le même jeu de données, plusieurs paramètres génétiques des populations Deli et La Mé susceptibles d'expliquer les résultats d'A. Nyouma : persistance des phases et du DL, fréquence de l'allèle de référence, pourcentage d'homozygotie et effets des SNP estimés par RRBLUP et LASSO Bayésien. Ce travail a mis en évidence que la densité de marquage utilisée (7 324 SNP) était suffisante pour atteindre une relativement bonne persistance des phases et du DL entre Deli et La Mé pour des SNP adjacents. Ainsi, plus de 40% des haploblocs de taille <3,600 bp ou <0.20 cM étaient communs aux deux populations (Figure 24), et la corrélation entre les deux populations dans les valeurs de LD (*r*) calculées entre paires de SNP étaient >0.6 pour des SNP séparés par des distances <0.5 cM ou <1 kbp. Ces ressemblances entre Deli et La Mé peuvent s'expliquer par le fait que la population Deli a des ancêtres africains, plantés en 1848 en Indonésie (voir 2.2.4.e), soit avec un nombre de générations depuis la divergence entre les deux populations probablement bien plus faible que les valeurs indiquées par Ibánez-Escriche et al. (2009) et Esfandyari et al. (2015). Cependant, le travail d'E. Seyum a aussi montré une absence de corrélation dans les effets des marqueurs pour tous les caractères (non publié, voir l'exemple pour PM sur la Figure 25). Des résultats similaires ont été obtenus sur le maïs, avec une bonne conservation des phases entre groupes hétérotiques Flint et Dent mais une absence de corrélation dans les effets des marqueurs (Technow et al., 2014). Deux hypothèses peuvent être envisagées pour expliquer ces résultats :

(i) la persistance des phases entre SNP n'est pas forcément un bon indicateur de la persistance des phases entre SNP et QTL. Des QTL peuvent par exemple présenter des polymorphismes causaux récents, et/ou des polymorphismes causaux qui ne seraient pas biallèliques comme les SNP. Dans le cas d'une faible persistance des phases entre SNP et QTL, les effets aux marqueurs seront faiblement corrélés entre les populations parentales, mais on s'attendrait alors à ce que le modèle PSAM donne de meilleures précisions que ASGM. Pourtant, ce n'est pas le cas dans nos études (Nyouma et al., 2022 ; Nyouma et al., 2020), ni dans celles sur le maïs (Technow et al., 2014). L'explication serait que l'essentiel de la précision de la SG dépend de la capacité des marqueurs à capturer l'apparentement entre les individus de calibration et les candidats à la sélection, ce qui ne nécessite pas forcément une association physique entre allèles aux SNP et aux QTL (Technow et al., 2014).

(ii) la persistance observée des phases entre Deli et La Mé au niveau des SNP pourrait effectivement traduire la persistance des phases entre SNP et QTL, et PSAM serait donc en théorie un modèle pertinent. Cependant, dans la pratique, PSAM pourrait s'avérer moins performant sous l'effet d'une population de calibration de taille insuffisante car, compte tenu de la décomposition des valeurs additives des hybrides en deux effets parentaux, PSAM a plus de paramètres à estimer qu'ASGM. Il serait intéressant d'étudier cet aspect plus en détail chez le palmier à huile, ce qui pourra se faire dans un premier temps par simulations.

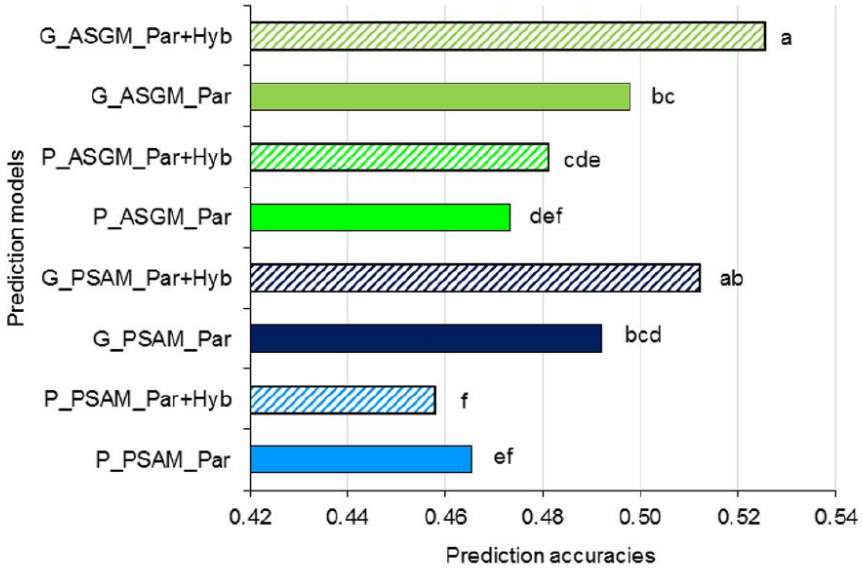


Fig. 3 Average prediction accuracies of prediction models across traits. Values with the same letter are not significantly different at $P = 5\%$

Figure 23 Précisions de prédictions moyennes sur neuf composantes du rendement en huile de palme dans une validation entre dispositifs expérimentaux, en fonction du modèle (Nyouma et al., 2022).

Types d'approches : « G_ » : modèle génomique, « P_ » : modèle témoin généalogique, « PSAM » : effets additifs des croisements décomposés en un effet par population parental, « ASGM » : pas de décomposition, « _Par » : seules les données moléculaires (ou généalogiques) des parents sont utilisées dans le modèle, « _Par+Hyb » : utilisation des données moléculaires (ou généalogiques) des parents et des individus hybrides. « P_PSAM_Par » correspond au modèle de Stuber et Cockerham (1966).

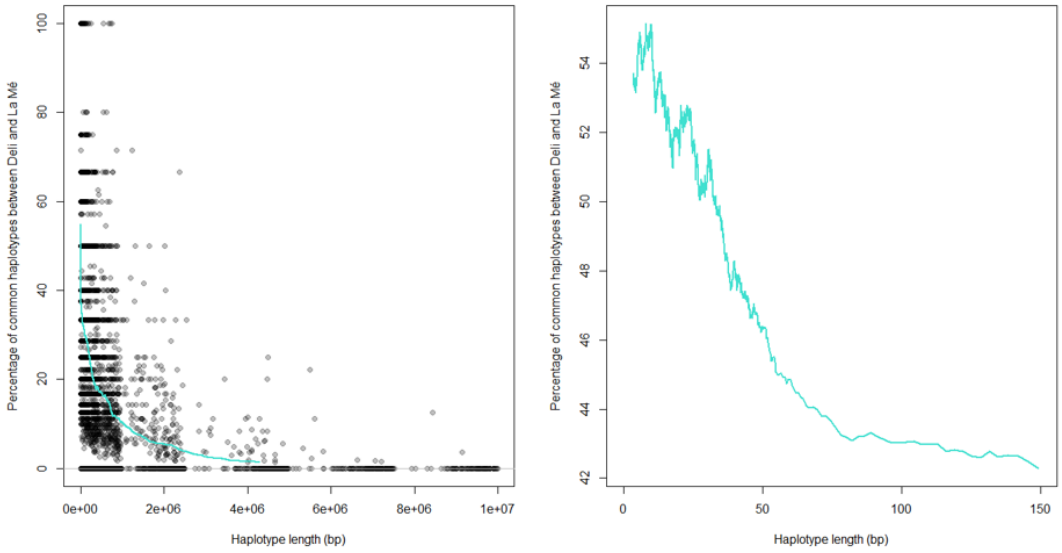


Figure 11: Percentage of common haplotypes between Deli and La Mé oil palm breeding populations according to the haplotype length in bp. Each dot represents a haplotype. Color intensity indicates density of overlapping dots. The smoothing curve in turquoise is the rolling average.

Figure 24 Persistance des phases au niveau des SNP entre les populations Deli et La Mé (Seyum et al., 2022b)

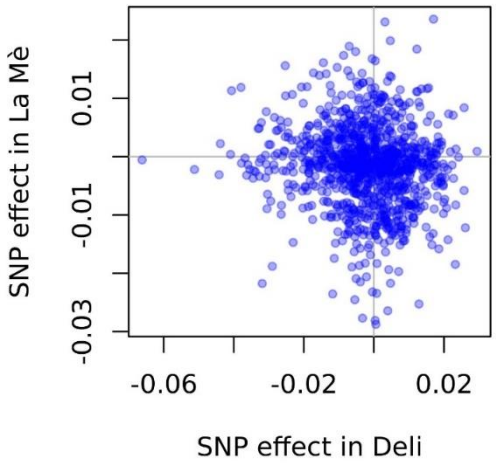


Figure 25 Corrélation dans les effets aux marqueurs entre Deli et La Mé pour le caractère poids moyen des régimes avec un modèle génomique PSAM (unpublished)

d. Modèles multi-caractères

Lorsqu'on considère des caractères avec un niveau de corrélation suffisamment élevé mais avec des niveaux d'héritabilité contrastés, l'utilisation de modèles multi-caractères peut augmenter la précision de la prédiction pour les caractères à faible héritabilité (Tong et Nikoloski, 2021). Chez le palmier à huile, le travail conduit lors du stage de M2 d'Alexandre Marchal a confirmé ce résultat sur le nombre de régimes et leur poids moyen (Marchal et al., 2016). Les modèles multi-caractères offrent donc la possibilité d'améliorer la précision des prédictions sans coût supplémentaire (hormis l'augmentation du besoin en ressources informatiques), et ils devraient donc être systématiquement évalués lorsque des corrélations existent entre les caractères d'intérêt, ou entre les caractères d'intérêt et des caractères secondaires Chez le palmier à huile, d'autres caractères peuvent être considérés, comme le pourcentage de pulpe et d'amande dans les fruits, fortement corrélés. Chez l'hévéa, le rendement en latex et la circonférence du tronc se prêteraient aussi à ce genre d'approche.

2.3.2. Déséquilibre de liaison (DL) et taille efficace (Ne)

Le déséquilibre de liaison (DL) entre marqueurs et QTL et la taille efficace (*Ne*) ont des effets interdépendants qui influencent fortement la précision de la SG (Heffner et al., 2009 ; Isik, 2014 ; Lebedev et al., 2020). Le DL est défini comme l'association non aléatoire d'allèles à deux loci ou plus dans des haplotypes (Slatkin, 2008 ; Weir, 1979). Le DL entre deux loci est mesuré sur la base de la fréquence des allèles, en utilisant des indices tels que r^2 et D (voir Annexe 7 pour plus de détails sur le calcul du DL). Une hypothèse clé de la SG est que la couverture dense du génome permet que chaque QTL contrôlant le caractère d'intérêt soit en DL avec au moins un marqueur. Le DL est donc l'un des facteurs majeurs affectant le nombre de marqueurs nécessaires à la SG (Heffner et al., 2009 ; Lebedev et al., 2020). Une valeur de r^2 entre marqueurs adjacents de 0,3 est considérée comme un minimum pour obtenir des résultats fiables en prédictions génomiques et dans les études de GWAS (Bejarano et al., 2018). Une bonne connaissance de ce paramètre dans la population cible est donc particulièrement intéressante pour définir la densité de marqueurs requise pour la SG. Il est donc utile d'explorer les événements historiques, tels que les goulots d'étranglement, la dérive génétique, la sélection naturelle

et artificielle, qui ont pu façonner le profil de DL (Flint-Garcia et al., 2003 ; Gupta et al., 2005 ; Hamilton, 2021 ; Mackay et Powell, 2007 ; Rogers, 2014 ; Slatkin, 2008).

Le profil de DL est largement déterminé par *Ne*, qui est la taille d'une population idéale de Wright-Fisher qui donnerait lieu au même degré de dérive génétique aléatoire que la population réelle (Caballero, 1994). Il existe une relation inverse entre *Ne* et DL, les taux élevés de dérive génétique et de consanguinité dans les populations à faible *Ne* entraînant un fort DL entre les marqueurs et les QTL par rapport aux populations à *Ne* élevé (Grattapaglia, 2014 ; Lin et al., 2014 ; Thistlethwaite et al., 2020). Lorsque *Ne* diminue et que le DL augmente, les individus ont tendance à partager des haplotypes plus longs, ce qui permet d'atteindre une bonne précision de prédiction génomique (Clark et al., 2012 ; Heffner et al., 2009 ; Isik, 2014 ; Lebedev et al., 2020). Pour une densité de marqueurs, une taille de population de calibration et un caractère donnés, le DL et la précision de la SG sont plus élevés dans les populations à faible *Ne* que dans les populations à *Ne* élevée (Grattapaglia, 2014 ; Lin et al., 2014 ; Solberg et al., 2008).

J'ai étudié le DL et *Ne* chez le palmier à huile dans la perspective de mieux comprendre les résultats obtenus en termes de précisions de SG, en lien avec la densité de marquage. J'ai fait de premières estimations de *Ne* lors mon doctorat (Cros et al., 2014), puis j'ai planifié une étude plus complète sur les Deli et La Mé dans le cadre du doctorat d'Essubalew Seyum (Seyum et al., 2022b). Les résultats ont montré que la vitesse et l'amplitude de la décroissance du DL en fonction des distances séparant les SNP différait entre les populations (Figure 26). Le DL plus élevé chez les Deli provient du plus grand nombre de générations de sélection, de consanguinité et de dérive génétique que les La Mé, ainsi que d'un goulot d'étranglement plus marqué (voir 2.2.4.e). Quand on considère les distances génétiques, le r^2 atteint 0,3 pour des SNP séparés par environ 220 kbp chez les Deli et 210 kbp chez les La Mé (Figure 26). Un ensemble de 10 000 SNP couvrant l'ensemble du génome (1,8 Gbp) serait suffisant pour atteindre un r^2 de 0,3 entre SNP adjacents dans les deux populations étudiées. Ce chiffre est en accord avec les valeurs obtenues en évaluant l'effet du nombre de SNP sur la précision de SG (0).

Pour le calcul des valeurs de *Ne* dans les principales populations considérées dans mon travail (Deli, La Mé et Yangambi), j'ai utilisé une approche basée sur le pédigrée (Gutiérrez et Goyache, 2005) et une sur le DL (Waples et Do, 2008). Sans surprise compte tenu de leur histoire, ces populations ont des tailles efficaces faibles (≤10) (Cros et al., 2014 ; Seyum et al., 2022b).

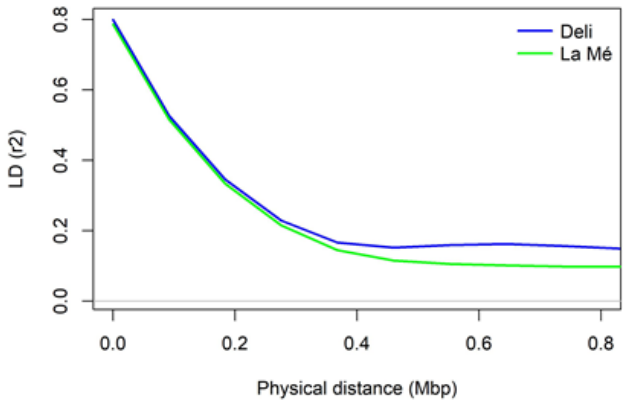


Figure 26: Profils de déséquilibre de liaison, mesuré par le r^2 entre SNP, en fonction de la distance génétique (gauche) et physique (droite) séparant les SNP, pour des individus utilisés dans le programme d'amélioration génétique de PalmElit appartenant aux populations Deli et La Mé (Seyum et al., 2022b).

2.3.3. Marqueurs moléculaires

a. Type de marqueurs

Quand j'ai commencé à travailler sur la SG du palmier à huile puis de l'hévéa, les seules données de marquage disponibles étaient pour des marqueurs SSR. J'ai donc utilisé ces données pour les premières évaluations (Cros et al., 2015b ; Cros et al., 2019 ; Marchal et al., 2016). Dans des situations avec un DL fort, comme dans les populations d'amélioration du palmier à huile et dans une famille de plein-frères comme pour l'hévéa, les SSR peuvent permettre d'atteindre une densité de marquage suffisante pour des prédictions génomiques. Par contre, ce type de marqueurs n'est pas adapté pour une application en routine de la SG, dont le potentiel en termes d'augmentation de l'intensité de sélection (voir 2.2.3.e) est directement lié au nombre d'individus qui seront génotypés. Dans les projets qui ont suivi, nous avons donc opté avec nos partenaires, PalmElit et IFC, pour l'utilisation d'une méthode de génotypage haut-débit, le GBS (2.2.3.d) (Cros et al., 2017 ; Munyengwa et al., 2021 ; Nyouma et al., 2022 ; Nyouma et al., 2020).

Le GBS génère des quantités relativement importantes de données manquantes, qui doivent être imputées. Par exemple, dans les études de Munyengwa et al. (2021) et Nyouma et al. (2020), le pourcentage de données manquantes variait de 11% à 23,08%, pour des jeux de données comprenant entre 3 420 et 15 054 SNP. La méthode d'imputation est donc un point important pour la précision de la SG avec du génotypage par GBS. Un grand nombre de méthodes ayant été développées récemment, des comparaisons sur le palmier à huile et l'hévéa me sont apparues utiles. J'ai encadré Norman Munyengwa en stage de Master 2 pour faire ce travail sur l'hévéa. Il a comparé 10 méthodes qui ont été évaluées sur le pourcentage de génotypes correctement imputés, et sur la précision de prédictions génomiques sur deux jeux de données comprenant au total trois caractères. Il a montré que le logiciel Beagle (Browning et al., 2018), très largement utilisé, était particulièrement efficace, mais que LinkImputeR (Money et al., 2017) et FImpute (Sargolzaei et al., 2014) étaient des alternatives intéressantes. En particulier, LinkImputeR donnait le plus fort pourcentage de génotypes correctement imputés (Figure 27) et de bonnes précisions (Munyengwa et al., 2021). Par ailleurs, il n'a pas besoin de la position des marqueurs pour faire l'imputation, ce qui représente un avantage certain pour les espèces n'ayant pas encore de génome de référence assemblé en pseudo-molécules. J'ai ensuite conduit un travail complémentaire sur le palmier à huile, qui me semblait intéressant compte tenu de la différence de type de population (familles de plein-frères chez l'hévéa, population complexe chez le palmier à huile). J'ai comparé sept méthodes d'imputation, en termes de précision de SG entre dispositifs expérimentaux pour 10 caractères d'intérêt (non publié). Cette étude a, à nouveau, montré la bonne performance de LinkImputeR, qui a donné des précisions légèrement supérieures en moyennes (Figure 28), bien que les différences significatives, recherchées par caractères, aient été très peu nombreuses. Les bons résultats obtenus par cet outil dans deux études sur des jeux de données contrastés pousse quand même à recommander cette méthode. Par contre, elle ne phase pas les données moléculaires, contrairement à Beagle. Dans le cas où la reconstruction des phases est nécessaire, Beagle est alors à privilégier.

Sur le palmier à huile, d'autres groupes ont utilisé des puces à SNP (Ithnin et al., 2017 ; Kwong et al., 2017a ; Kwong et al., 2017b). Suite au développement de puces à SNP dans des projets portés au Cirad par Norbert Billotte et Sébastien Tisné, j'ai fait une comparaison de la précision de prédictions génomiques réalisées à partir de données moléculaires obtenues par GBS et par une puce à SNP Axiom® 55K (non publié). Ce travail a montré que les deux méthodes de génotypage donnaient les mêmes précisions de SG (avec peut-être un léger avantage à la puce dans certains cas, mais la

différence n'était pas-significative) (Figure 29). Dans mes recherches, j'utilise désormais le génotypage par puce, car il présente aussi les avantages suivants par rapport au GBS :

(i) une plus grande densité de marquage : le GBS donne des densités suffisantes pour la SG mais pour d'autres approches, notamment la détection de QTL, il est intéressant d'avoir plus de marqueurs. Les génotypages par puce permettent de produire des jeux de SNP qui peuvent plus facilement servir à différentes études.

(ii) des données moléculaires de plus grande qualité, avec moins de données manquantes et d'erreur de génotypage, ce qui facilite le travail.

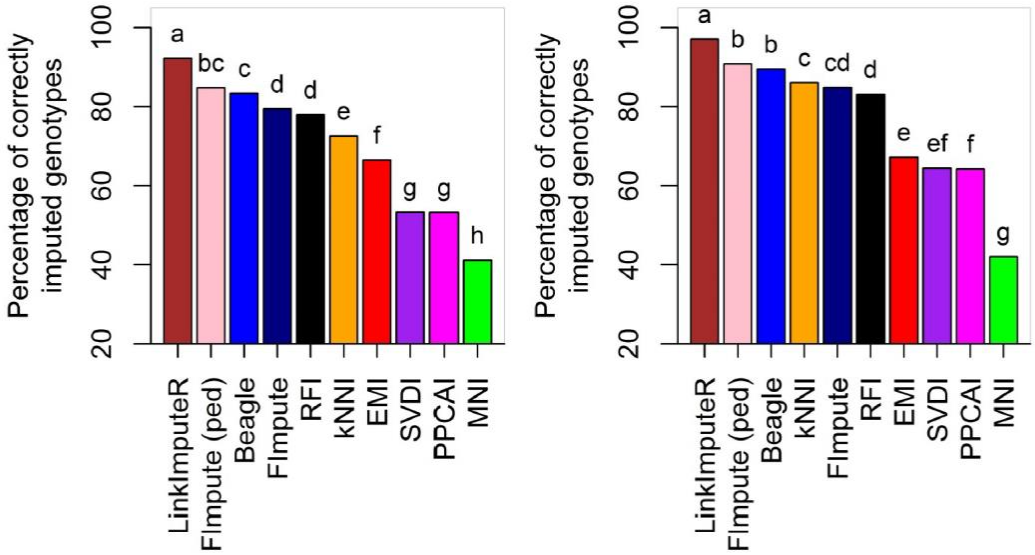
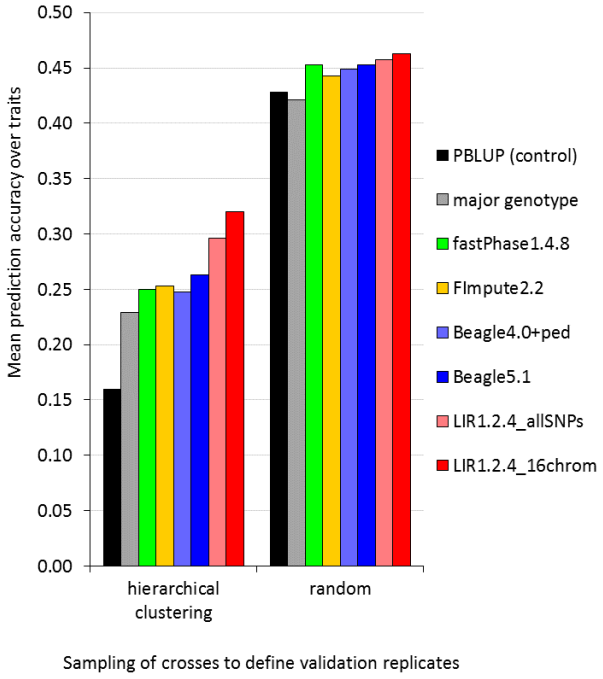


Fig. 4. Percentage of correctly imputed genotypes according to imputation method in dataset 1 (left) and dataset 2 (right). Figures are means over three replicates. Values with the same letter are not significantly different within a dataset at $P = 1\%$.

Figure 27 Pourcentage de génotypes de GBS correctement imputés dans une famille de plein-frères de l'hévéa, en fonction de la méthode d'imputation (Munyengwa et al., 2021)



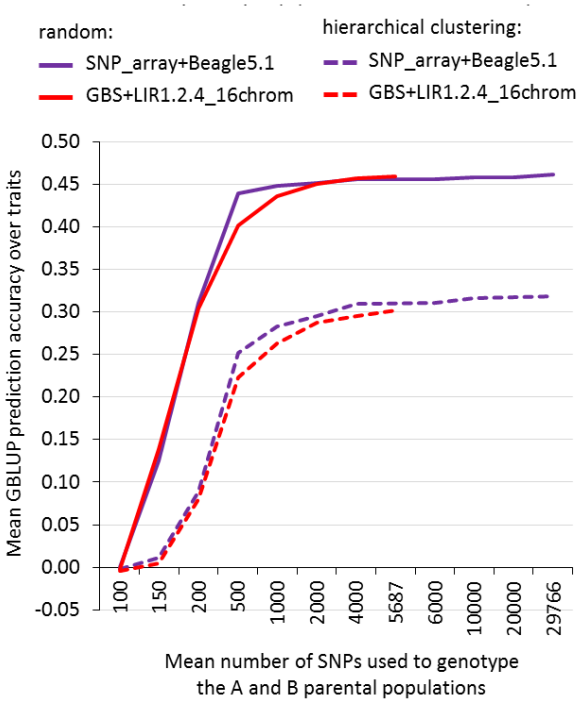


Figure 28 Précision de sélection moyenne sur 10 caractères du palmier à huile pour des prédictions génomiques réalisées entre dispositifs expérimentaux, en fonction de la méthode d'imputation des génotypes de GBS sporadiques manquants et de l'apparentement entre calibration et validation (« random » fort, « clustering » faible) (Cros et al., non publié).

Figure 29 Précision de sélection moyenne sur 10 caractères du palmier à huile pour des prédictions génomiques réalisées entre dispositifs expérimentaux, en fonction de la méthode de génotypage, de la densité de marquage et de l'apparentement entre calibration et validation (« random » fort, « clustering » faible) (Cros et al., non publié).

b. Densité de marquage

La densité de marquage nécessaire pour la SG dépend du type de marqueurs, du mode d'échantillonnage des marqueurs, du caractère et, comme on l'a vu précédemment, de la population (DL/*Ne*). Augmenter le nombre de marqueurs améliore la précision de la SG, jusqu'à ce qu'un plateau soit atteint (Isik, 2014 ; Lin et al., 2014 ; Meuwissen et al., 2001 ; Robertsen et al., 2019 ; Solberg et al., 2008).

Chez le palmier à huile, cet aspect a été étudié dans mes recherches et dans les travaux de deux étudiants. Les résultats ont montré que la précision de la SG plafonnait à partir de 100 à 200 SSRs (Marchal et al., 2016) et 500 à 7 000 SNP de GBS, en fonction du caractère (voir Figure 30 pour l'exemple de la production de régimes) (Cros et al., 2017 ; Nyouma et al., 2020). Avec des SNP de puce, la densité de marquage nécessaire est semblable à celles avec des SNP de GBS (Figure 29). Un travail similaire a été réalisé chez l'hévéa avec des stagiaires de M2, et nous avons mis en évidence que 300 SSR étaient suffisants pour atteindre la précision maximale pour la production de latex dans la famille de plein-frère considérée (Figure 31) (Cros et al., 2019), ou 1 600 SNP de GBS (Munyengwa et al., 2021). Ces densités de marquage sont faibles par rapport à ce qui est généralement utilisé dans d'autres espèces, mais cela résulte du fort DL de nos populations, soit dû à leur histoire (palmier à huile), soit par construction (utilisation de familles de plein-frères chez l'hévéa).

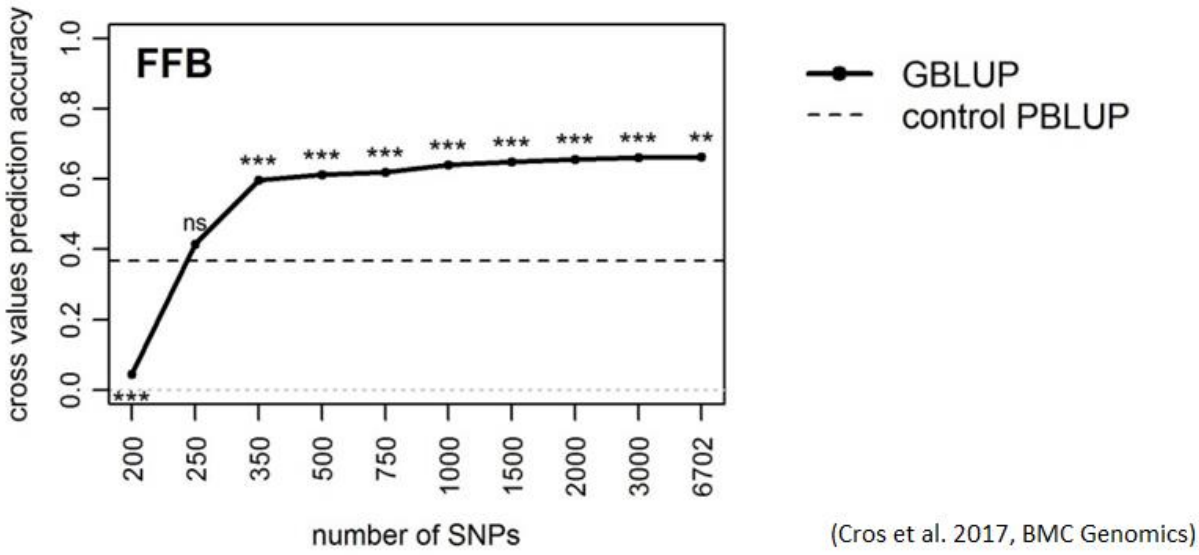


Figure 30 Précision de la SG entre dispositifs expérimentaux pour la prédiction du poids total de régimes (FFB) chez des croisements hybrides de palmier à huile, en fonction de la densité de marquage SNP (Cros et al., 2017)

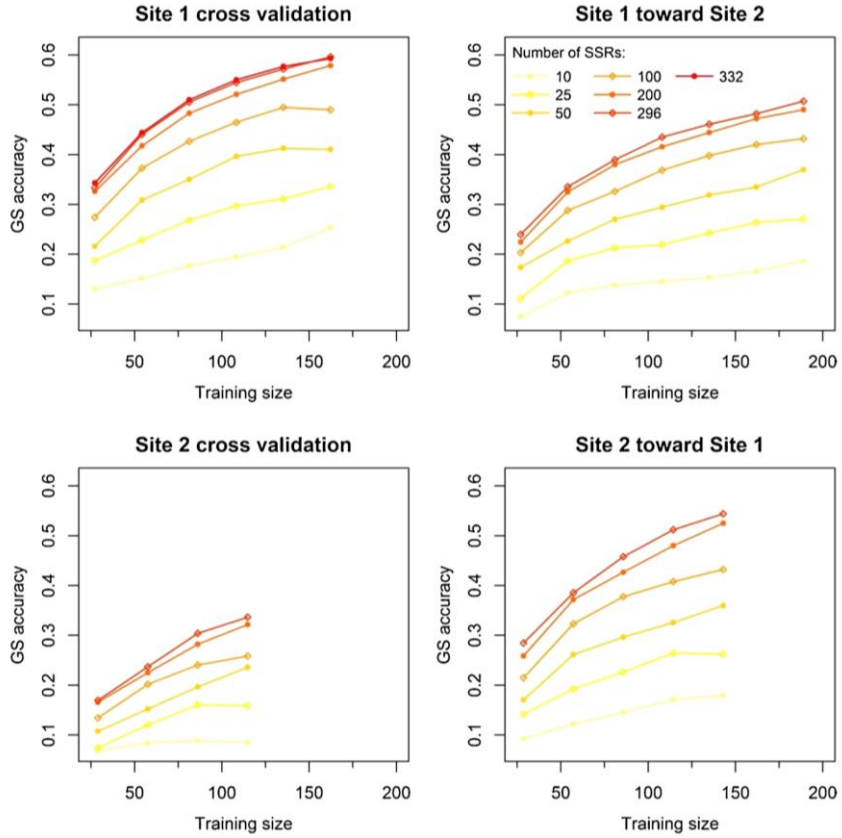


Fig. 3. GS accuracy in predicting rubber yield according to number of clones used to train the GS prediction model (training size), number of SSRs, and validation approach. For a given number of SSRs, random SNPs were sampled. Values are means of seven to 1400 replicates, depending on training size, number of SSRs, and validation approach.

Figure 31 Précision de la SG pour la prédiction de la production de latex chez des clones d'une même famille de plein-frères d'hévéa, en fonction de la densité de marquage SSR, de la taille de la population de calibration et de la méthode de validation (validation croisée intra-site ou validation inter-site) (Cros et al., 2019)

c. Sélection des SNP

La sélection des SNP peut réduire le coût de la SG et/ou en augmenter la précision. Avec des SNP de GBS, des filtres évidents peuvent être appliqués sur les SNP en fonction de leur profondeur de séquençage moyenne ou sur leur pourcentage de données manquantes. Chez le palmier à huile, l'utilisation des SNP de GBS avec le moins de données manquantes a augmenté significativement la précision des prédictions par rapport à l'utilisation de tous les SNP ou d'un nombre équivalent de SNP choisis au hasard (Cros et al., 2017). Ceci montre que l'imputation des génotypes manquants peut introduire des erreurs. Ce résultat n'a par contre été observé que sur un seul caractère (Figure 32). Par ailleurs, depuis cette étude, les progrès réalisés en termes de génotypage et d'imputation laissent penser que cet aspect ne pose désormais plus vraiment de problèmes. En effet, les génotypages par puce permettent d'avoir une proportion insignifiante de données manquantes. Dans le cas où l'on utilise toujours le GBS, les profondeurs que l'on peut obtenir actuellement, combinées à des méthodes d'imputation performantes, permettent d'avoir les mêmes précisions qu'avec une puce à SNP (voir 2.3.3.a).

Chez l'hévéa, avec du GBS, nous avons constaté que les précisions de SG étaient plus élevées avec le sous-ensemble de marqueurs qui avaient pu être positionnés sur une carte génétique par rapport à l'ensemble des marqueurs, de façon quasi-systématique quel que soit le caractère et la méthode d'imputation (+4,3% en moyenne) (Munyengwa et al., 2021). Cela suggère que le processus de cartographie génétique conduit à filtrer les SNP sur des critères pertinents pour les prédictions génomiques, tels que la cohérence entre les génotypes parentaux et la ségrégation dans la descendance (qui est liée à un faible taux d'erreurs de génotypage et à une distorsion réduite de la ségrégation) et l'élimination des marqueurs redondants (situés à des positions similaires sur la carte génétique, et qui fournissent probablement des informations génomiques similaires en raison du DL).

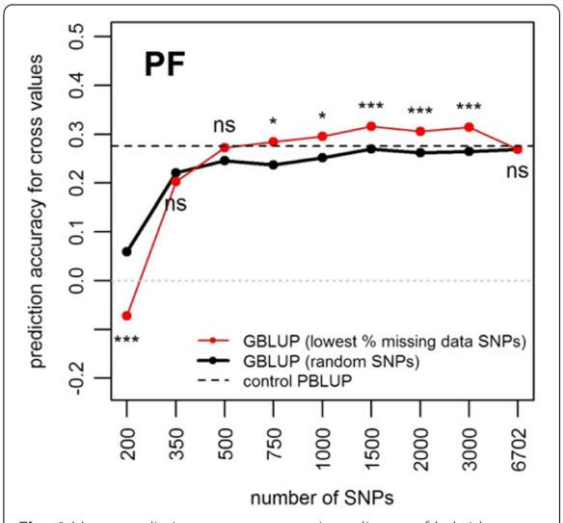


Fig. 4 Mean prediction accuracy over six replicates of hybrid crosses, when predicting cross values of Site 2 using a model calibrated on Site 1 for the PF trait, using GBLUP with random subsets of SNPs (solid black line), SNP subsets minimizing the percentage of missing data (red line) and control PBLUP (dashed line). The last level in the number of SNPs (6702) is the mean number of SNPs in Group A and Group B. Significance of ANOVA for method of SNP sampling: *** $P < 0.001$, * $0.01 \leq P < 0.05$, *ns* not significant. Values are means over $n = 26$ replicates of random marker subsets, $n = 3$ when minimizing percentage of missing data and $n = 1$ when using all the markers

Figure 32 Précision de la SG pour la prédiction des performances de croisements hybrides de palmier à huile entre dispositifs expérimentaux pour le pourcentage de pulpe dans les fruits (PF), en fonction du nombre de SNP et de la méthode de choix des SNP (aléatoire ou avec le moins de données manquantes) (Cros et al., 2017)

2.3.4. Populations de calibration et de validation

a. Apparentement entre populations de calibration et de validation

La précision de la SG est positivement corrélée avec le niveau d'apparentement entre populations de calibration et de sélection. Ceci tient au fait que, lorsque des individus sont fortement apparentés, ils ont en commun de longs haploblocs (Daetwyler et al., 2013 ; Isidro y Sánchez et Akdemir, 2021 ; Pszczola et al., 2012 ; Wientjes et al., 2013). J'ai ainsi observé, dans une validation croisée réalisée durant mon doctorat, une corrélation significative entre la précision de la SG et l'apparentement entre populations de calibration et de validation sur plusieurs composantes du rendement du palmier à huile (voir l'exemple pour un caractère sur la Figure 33) (Cros et al., 2015b). De la même façon, dans une validation entre dispositifs expérimentaux, j'ai observé que la précision de prédiction de la SG était largement réduite lorsque la population de validation était découpée en répétitions obtenues de façon à maximiser l'apparentement au sein des répétitions, ce qui réduisait aussi l'apparentement moyen avec la population de calibration, par rapport à un découpage aléatoire (Figure 28 et Figure 29).

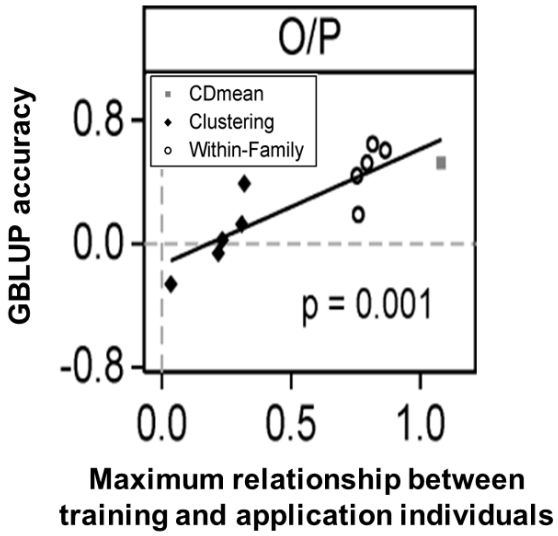


Figure 33 Précision de la SG en fonction de l'apparentement entre les populations de calibration et de sélection chez le palmier à huile pour le caractère pourcentage d'huile dans la pulpe (Cros et al., 2015b)

b. Taille et structure de la population de calibration

La taille de la population de calibration est l'un des facteurs les plus importants qui déterminent la précision de la SG, et plusieurs études ont montré que l'augmentation de la taille de la population de calibration améliorait la précision de la SG (Calleja-Rodriguez et al., 2020 ; Cericola et al., 2018 ; Combs et Bernardo, 2013 ; Isidro et al., 2015 ; Liu et al., 2018 ; Nielsen et al., 2016 ; Tan et al., 2017).

Nous avons observé cette augmentation pour la production de latex chez l'hévéa, avec un plateau en termes de précision atteint avec 200 individus dans notre famille de plein-frères (Figure 31) (Cros et al., 2019). Chez le palmier à huile, j'ai utilisé deux options pour augmenter la taille de la population de calibration :

(i) L'utilisation des données moléculaires des individus hybrides, en complément des données moléculaires des parents, qui étaient initialement les seules disponibles (voir 2.3.1). Cette méthode a été étudiée dans les simulations de mon doctorat puis sur des

données réelles par Achille Nyouma. Celui-ci a montré que l'ajout des données moléculaires de 399 individus hybrides pour la calibration augmentait la précision de la SG de 5% en moyenne sur les différents caractères (Figure 23) (Nyouma et al., 2022). La mise en œuvre de cette méthode a nécessité de combiner des apparentements génomiques et des apparentements généalogiques, ce qui peut se faire de façon simple avec la méthodologie du *single-step* GBLUP (Lourenco et al., 2020). Nous avons utilisé le *single-step* GBLUP dans plusieurs études, soit pour tenir compte du génotype d'un échantillon d'individus hybrides (Cros et al., 2015a ; Nyouma et al., 2022), soit pour ne pas exclure des parents sans données moléculaires (voir par exemple Cros et al., 2017). Avec les résultats des simulations (Cros et al., 2015a) et ceux du projet OPGP (non publié), il ressort que le génotypage d'un échantillon d'un millier d'individus hybrides répartis dans l'ensemble des croisements est une bonne stratégie pour le palmier à huile. L'utilisation de génotypes d'hybrides pour la SG chez cette espèce a aussi été envisagée par Kwong et al. (2017a). Ils ont travaillé sur une population de calibration de 1 218 individus mais n'ont pas fait varier la taille pour étudier son effet.

(ii) L'agrégation des données phénotypiques et moléculaires de cycles d'amélioration successifs. Ce point a fait l'objet du stage de Master 2 de Billy Tchounke. En simulant un schéma d'amélioration sur quatre cycles, il a montré que l'utilisation conjointe des données des deux premiers cycles pour calibrer le modèle de SG augmentait la réponse à la sélection de plus de 10% par cycle, en raison d'une plus grande précision de sélection, et ce malgré une réduction du niveau d'apparentement entre les populations de calibration et de sélection (Cros et al., 2018).

Les coûts de phénotypage sont une contrainte majeure en SG, et ce d'autant plus que les coûts de génotypage ont considérablement diminué (Akdemir et Isidro-Sánchez, 2019). Cette contrainte financière s'applique particulièrement aux cultures pérennes, car leurs évaluations phénotypiques nécessitent de grandes surfaces sur plusieurs années. La définition de populations de calibration optimisées est donc particulièrement intéressante dans ce contexte. L'optimisation de la population de calibration est le processus de sélection, au sein d'un ensemble d'individus qui pourrait être utilisé pour calibrer le modèle de SG, d'un échantillon d'individus qui prédit le mieux la valeur génétique des candidats à la sélection. Plusieurs méthodes ont été développées, notamment CDmean, PEVmean, l'échantillonnage stratifié et EthAcc (Isidro y Sánchez et Akdemir, 2021). Dans le cas du palmier à huile, j'ai testé le CDmean (Rincent et al., 2012). Le CDmean est applicable lorsque l'on dispose d'un ensemble d'individus génotypés et que l'on souhaite en tirer un sous ensemble qui sera phénotypé pour constituer une population de calibration avec laquelle on prédira la GEBV des individus restants. Le critère d'optimisation est le coefficient de détermination généralisé. Il est fonction de la variance des erreurs de prédiction (Annexe 3) et de la variance génétique, et correspond au carré de la corrélation attendue entre la vraie valeur et la valeur estimée du contraste des valeurs génétiques. Sur le palmier à huile, il s'est avéré efficace en donnant les précisions de SG les plus élevées, atteignant 0.79 en moyenne sur les différentes composantes du rendement (Cros et al., 2015b).

2.3.5. Héritabilité des caractères

La précision de la SG est affectée par l'héritabilité des caractères, avec des héritabilités élevées aboutissant à des précisions plus fortes (Hayes et al., 2009 ; Lin et al., 2014 ; Meuwissen et al., 2001).

Chez le palmier à huile, j'ai mis en évidence une corrélation positive entre h^2 et la précision de la SG sur huit composantes du rendement en huile de palme au sein du groupe B (Cros et al., 2015b).

2.4. Schéma d'amélioration génomique et rythme du progrès génétique

Je suis intéressé par l'étude des facteurs influençant la précision des prédictions génomiques du point de vue théorique, mais je considère cette partie de mon travail comme une étape intermédiaire, la finalité étant essentiellement d'aboutir à des conclusions qui permettent des prises de décisions pour restructurer les schémas conventionnels d'amélioration en versions génomiques plus efficaces. Sur le principe, la SG peut facilement s'intégrer aux schémas d'amélioration du palmier à huile ou de l'hévéa (voir par exemple pour le palmier à huile Figure 34). Cependant, dans le détail, plusieurs options sont possibles, et des études sont donc nécessaires pour identifier la meilleure, et s'assurer qu'elle aboutisse à un progrès génétique plus important que le schéma conventionnel.

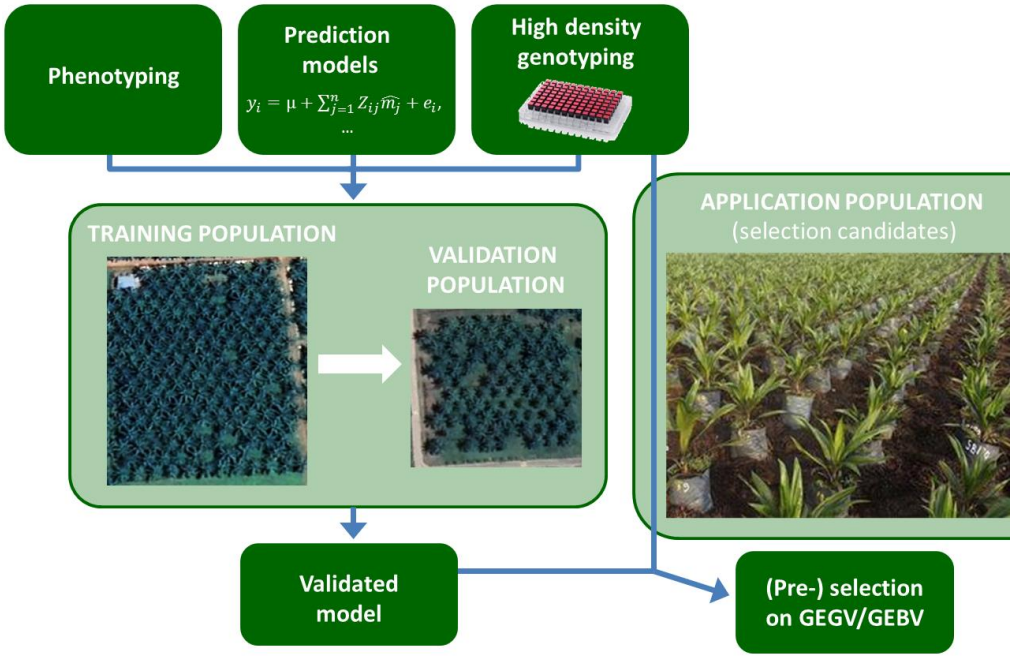


Figure 34 Principe de la sélection génomique appliquée au palmier à huile (Cros et al, 2019, PIPOC)

2.4.1. Rythme du progrès génétique

a. Chez le palmier à huile

La SRR phénotypique conventionnelle et la SRR génomique ont été comparées pour l'amélioration du poids total de régimes sur quatre cycles (Cros et al., 2015a). Le schéma de SG incluait des cycles avec des tests sur descendance hybride et des cycles de sélection purement génomique. Les tests sur descendance étaient utilisés pour calibrer un modèle de SG appliqué pour effectuer une sélection parmi les individus Deli et La Mé non évalués du même cycle et/ou des générations suivantes. La simulation a quantifié l'effet de trois paramètres sur la réponse annuelle à la sélection : la fréquence des tests de descendance (à chaque cycle, uniquement au cycle 1, ou aux cycles 1 et 3), le nombre

d'individus soumis à la SG (120, comme dans la RRS, et 300) et la stratégie de SG (génotypage limité aux parents des hybrides de calibration [RRGS_PAR] ou avec le génotypage des individus hybrides [RRGS_HYB], voir 2.3.1.c). Ce travail a montré que la SG pouvait augmenter le progrès génétique annuel en réduisant l'intervalle moyen de générations et en augmentant l'intensité de la sélection, malgré le fait que la précision de la SG pour les individus non évalués était inférieure à celle des individus testés en croisements hybrides (Figure 36). Parmi les stratégies évaluées, la stratégie RRGS_HYB avec le génotypage de 1 700 individus hybrides, la calibration du modèle uniquement à la première génération et 300 candidats à la sélection par population et par génération était la plus efficace, conduisant à un progrès génétique annuel supérieur de 72% à celui de la SRR (Figure 35). En outre, RRGS_PAR, avec une calibration du modèle toutes les deux générations et 300 candidats à la sélection, s'est révélée être une alternative intéressante car, bien que son progrès génétique soit plus faible (46% de plus que la SRR), elle présente une plus faible variabilité du progrès génétique, un coût réduit et une augmentation plus lente de la consanguinité au fil des cycles dans les populations parentales.

Le progrès génétique permis par la SG a aussi été évalué dans une étude conduite par un autre groupe (Wong et Bernardo, 2008). En simulant plusieurs cycles d'une sélection au sein d'une famille de plein-frères d'un des groupes hétérotiques, ils ont constaté que la SG et la sélection phénotypique surpassaient la SAM basée sur les QTL en termes de réponse à la sélection, tandis que la SG surpassait la sélection phénotypique lorsque la taille de la famille atteignait 50 à 70 individus, et augmentait alors la réponse à la sélection de 4 à 25 %, selon la taille de la famille, l'héritabilité et le nombre de QTL.

Ces simulations ont donc donné des résultats très prometteurs. Cependant, les études empiriques, même si elles ont montré que les précisions de la SG pouvaient être élevées, ont également révélé que la SG n'était pas efficace pour toutes les composantes du rendement. En effet, pour certains caractères, la SG a donné une faible précision (<0,2) et/ou n'a pas été capable de capturer la ségrégation mendélienne, c-à-d les variations génétiques au sein des familles de plein-frères de parents candidats à la sélection (Cros et al., 2017 ; Cros et al., 2015b ; Nyouma et al., 2022 ; Nyouma et al., 2020). Les simulations ont montré que le principal avantage de la SG était sa capacité à raccourcir les cycles de sélection en évitant les évaluations au champ dans certains cycles, mais cela n'est possible que si la SG est efficace pour toutes les composantes du rendement qui font actuellement l'objet d'une sélection phénotypique. Dans le cas contraire, les tests de descendance restent nécessaires dans tous les cycles de sélection. Par conséquent, l'application pratique actuellement envisagée pour commencer à mettre en œuvre la SG chez le palmier à huile est un schéma en deux étapes, avec une première étape de sélection génomique avant les tests de descendance. La SG serait donc utilisée pour améliorer la sélection phénotypique qui se fait actuellement avant les tests sur descendance sur les caractères les plus héritables, c-à-d essentiellement %PF et %OP (voir 2.2.4.g). La SG permettrait d'augmenter le nombre de caractères soumis à une sélection avant les essais hybrides, ce qui, à l'échelle du cycle d'amélioration, augmenterait l'intensité de sélection. Le potentiel d'une telle présélection génomique a été quantifié pour le caractère production de régimes, sur la base des précisions de SG obtenues empiriquement pour ce caractère (Cros et al., 2017). L'étude a montré que ceci améliorerait la performance des hybrides A×B sélectionnés de plus de 10 % lorsque 4 000 A et 4 000 B étaient génotypés par rapport à la méthode actuelle sans présélection sur ce caractère (Figure 37), grâce à une intensité de sélection plus élevée.

De la même façon, dans le cadre de son doctorat, A. Nyouma a montré qu'une étape de SG avant les essais clonaux améliorerait les performances des clones sélectionnés (Nyouma et al., 2020). En effet, pour la quasi-totalité des composantes du rendement considérées, les prédictions

génomiques étaient plus fortement corrélées aux valeurs génétiques clonales que les observations phénotypiques actuellement réalisées (Figure 38). Dans le cas où les valeurs clonales des individus de validation étaient prédites à partir de leurs génotypes et de leurs valeurs propres, la SG atteignait une précision de 0,53 en moyenne sur l'ensemble des caractères, contre 0,46 lorsque la prédiction était faite uniquement à partir de leurs génotypes. Ces résultats suggèrent deux possibilités en termes d'intégration des prédictions génomiques pour la sélection clonale du palmier à huile, qui rendront plus efficace la présélection d'ortets avant les essais clonaux et augmenteront l'intensité de sélection :

(i) SG à l'âge adulte sur toutes les composantes du rendement en utilisant conjointement les génotypes et les phénotypes des candidats à la sélection, réalisée au sein des meilleurs croisements effectivement évalués en essais hybrides

(ii) SG au stade pépinière sur un plus grand nombre de composantes du rendement que dans la sélection phénotypique actuelle, réalisée au sein d'une large population des meilleurs croisements possibles (identifiés sur la base des résultats des essais hybrides).

Une autres des conclusions des simulations que j'ai réalisées chez le palmier à huile était que la SG augmentait plus rapidement la consanguinité au sein des populations parentales. Pour pallier à ce problème, Billy Tchounke a considéré la *mate selection* (Gorjanc et Hickey, 2018 ; Kinghorn et al., 2010 ; Sánchez et al., 1999 ; Toro et Perez-Enciso, 1990), une méthode développée pour limiter l'augmentation de la consanguinité au fil des générations tout en optimisant le progrès génétique. Dans la méthode conventionnelle, les meilleurs individus sont sélectionnés car leur valeur génétique est supérieure à une certaine valeur (troncation), puis sont croisés entre eux de manière aléatoire. De son côté, la *mate selection* optimise la sélection et le plan de croisements entre individus sélectionnés. Dans le cadre de son doctorat, Billy Tchounke a implémenté la *mate selection* avec une optimisation par l'algorithme du recuit simulé. Son travail a montré que, par rapport à la méthode conventionnelle sans optimisation, la *mate selection* pouvait diminuer la consanguinité et augmenter le progrès génétique de façon significative, l'ampleur des effets dépendant de l'importance donnée par l'utilisateur à ces deux critères (Tchounke et al., under review). La solution optimale retenue par la *mate selection* différe de la solution conventionnelle par cinq caractéristiques : individus sélectionnés couvrant une plus large gamme de valeurs génétiques, réduction du nombre d'individus sélectionnés par famille, limitation des autofécondations, autofécondations effectuées de préférence sur les meilleurs individus, et nombre déséquilibré de croisements entre les individus sélectionnés, le nombre de croisements étant d'autant plus élevé que l'individu est meilleur. L'association de la SG et de la *mate selection* est donc apparue comme une stratégie particulièrement intéressante.

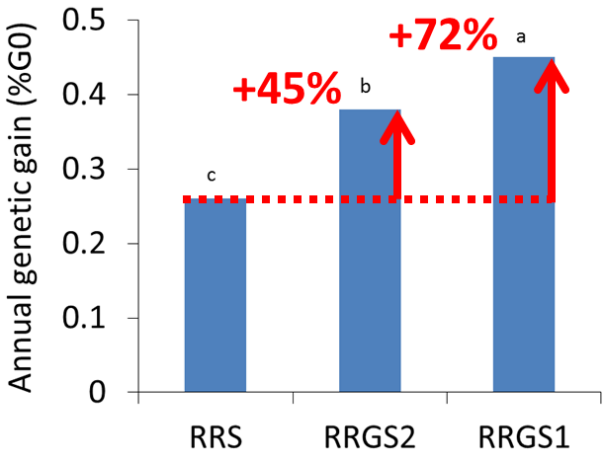


Figure 35 Progrès génétique estimé par simulation sur quatre cycles d'amélioration génétique chez le palmier à huile, en fonction du type de schéma d'amélioration

RRS, sélection récurrente réciproque phénotypique ; RRGS, sélection récurrente réciproque génomique avec : RRGS2 : calibration sur le génotype des parents des croisements et le phénotype des hybrides, 300 candidats à la sélection par cycle et groupe parental, tests sur descendance remplacés par de la SG dans les générations 2 et 4, RRGS1 : calibration sur le génotype des parents des croisements, le génotype d'individus hybrides et le phénotype des hybrides, 300 candidats à la sélection par cycle et groupe parental, tests sur descendance remplacés par de la SG dans les générations 2, 3 et 4 **(Cros et al., 2015a)**

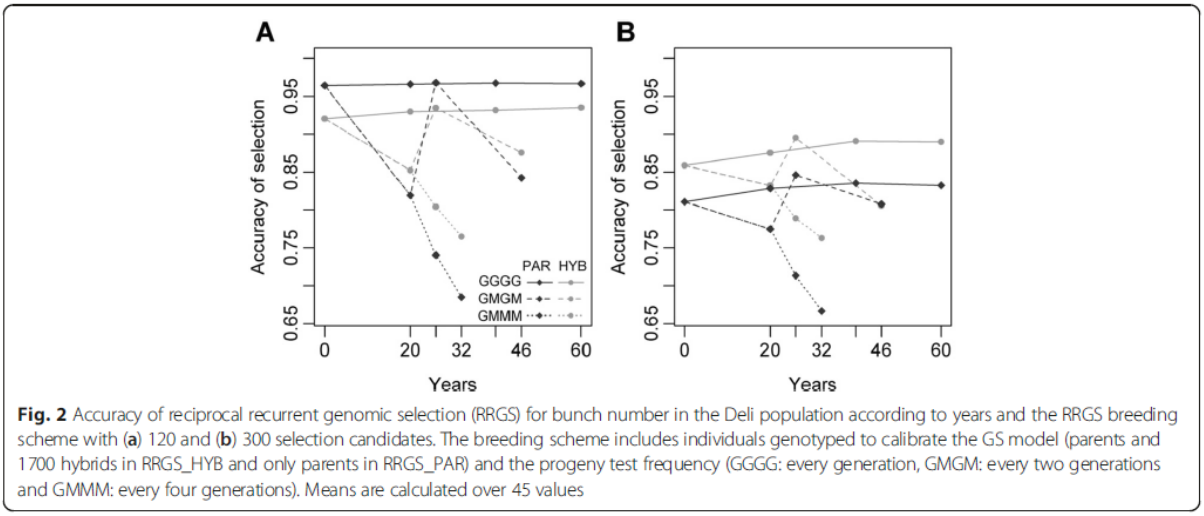


Fig. 2 Accuracy of reciprocal recurrent genomic selection (RRGS) for bunch number in the Deli population according to years and the RRGS breeding scheme with (**a**) 120 and (**b**) 300 selection candidates. The breeding scheme includes individuals genotyped to calibrate the GS model (parents and 1700 hybrids in RRGS_HYB and only parents in RRGS_PAR) and the progeny test frequency (GGGG: every generation, GMGM: every two generations and GMMM: every four generations). Means are calculated over 45 values

Figure 36 Précision de la SG pour le nombre de régimes en fonction des générations, de la fréquence de calibration du modèle et du modèle de prédiction (Cros et al., 2017)

(HYB : utilisation des données moléculaires des parents et des individus hybrides, PAR : utilisation des seules données moléculaires parentales, correspondant au modèle de Stuber et Cockerham (1966))

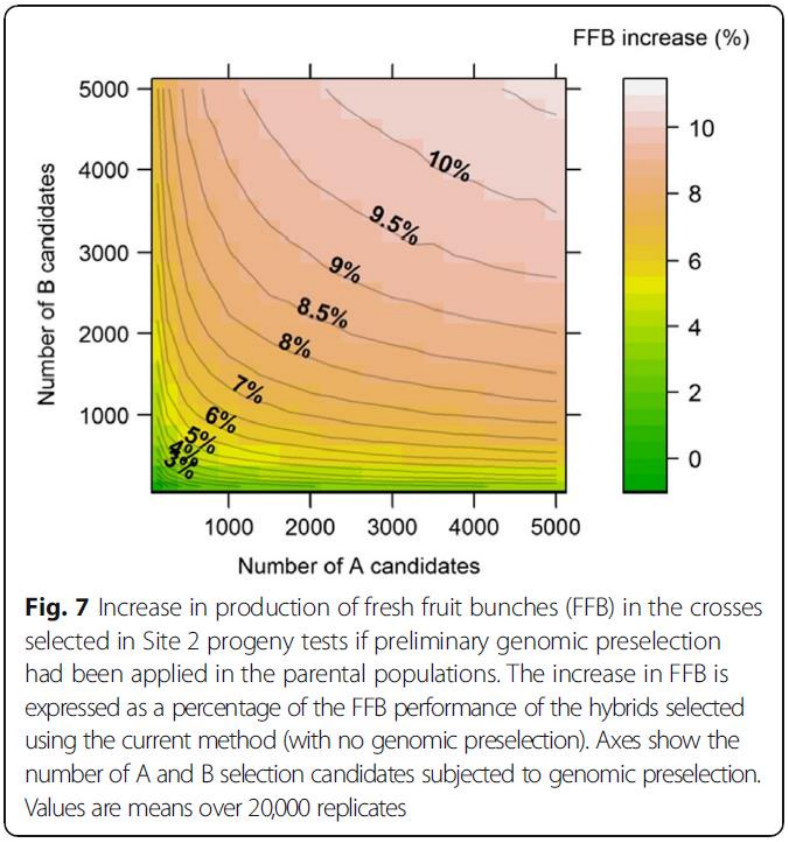


Fig. 7 Increase in production of fresh fruit bunches (FFB) in the crosses selected in Site 2 progeny tests if preliminary genomic preselection had been applied in the parental populations. The increase in FFB is expressed as a percentage of the FFB performance of the hybrids selected using the current method (with no genomic preselection). Axes show the number of A and B selection candidates subjected to genomic preselection. Values are means over 20,000 replicates

Figure 37 Augmentation de la production de régimes (FFB) avec la SG appliquée dans les populations parentales A et B avant les évaluations en croisements hybrides, exprimée en % du FFB des hybrides sélectionnés avec la méthode actuelle (sans SG) (Cros et al., 2017)

Les axes montrent le nombre de A et B soumis à la SG. Les valeurs sont des moyennes sur 20 0000 répétitions.

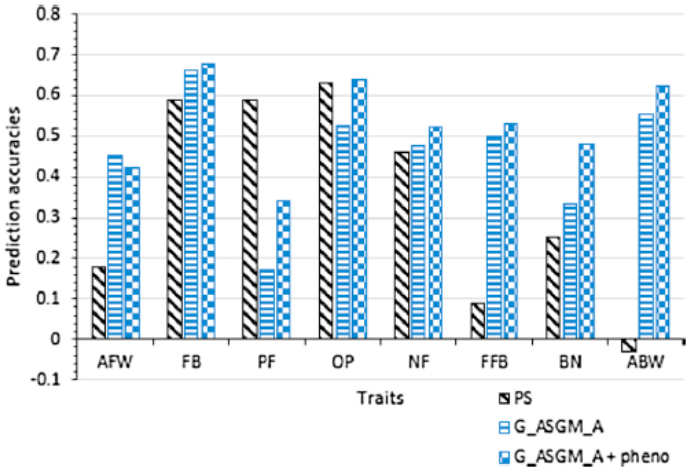


Fig. 4. Prediction accuracies of phenotypic selection (PS) and of the G_ASGM_A model without phenotypic data (G_ASGM_A) and with phenotypic data (G_ASGM_A + pheno) of ortets, on average over the best SNP datasets, and according to trait.

Figure 38 Précision de la sélection phénotypique (PS) et de la sélection génomique (G_ASGM) pour la prédiction des valeurs clonales (Nyouma et al., 2020).

G_ASGM_A : valeurs clonales des individus de validation prédites à partir de leurs seuls génotypes, G_ASGM_A + pheno : valeurs clonales des individus de validation prédites à partir de leurs génotypes et de leurs valeurs propres (Nyouma et al., 2020). La faible précision obtenue en SG pour le caractère PF est probablement due à la présélection phénotypique effectuée sur ce caractère avant les essais clonaux utilisés pour la validation (provoquant une non-normalité des valeurs de validation, préjudiciable au modèle de SG).

b. Chez l'hévéa

Le progrès génétique pour la production de latex chez l'hévéa que l'on pourrait obtenir par l'intégration de la SG a été étudié dans le cadre des stages de M2 de Luther Nkoulou et Jean Oum. Dans le schéma génomique envisagé (Figure 20, droite), l'étape SET est supprimée. A la place, la SG est appliquée sur des individus en pépinière avant de les évaluer en essai clonal (SSCT), avec un modèle calibré avec les données d'un premier SSCT. A partir d'une simulation calibrée avec les données réelles de la famille de plein-frères considérée, nous avons montré que le schéma génomique aboutissait à un progrès génétique plus important que le schéma phénotypique conventionnel, grâce à la plus grande précision de la SG par rapport à la sélection faite dans le SET (+56,7%) et à la plus grande intensité de sélection obtenue lorsque le nombre de candidats à la SG était suffisamment élevé (≥1000) (Figure 39). Ainsi, avec 3 000 individus soumis à la SG, le progrès génétique aurait augmenté de 10,3% (Cros et al., 2019).

Le progrès génétique permis par la SG chez l'hévéa a aussi été étudié par un autre groupe (Souza et al., 2019), qui a travaillé sur deux familles de plein-frères différentes et sur un autre caractère, le diamètre du tronc. Ils ont montré que la SG permettait d'atteindre un rythme du progrès génétique plus élevé que la sélection phénotypique conventionnelle, en considérant que la SG permettrait de raccourcir les cycles d'amélioration.

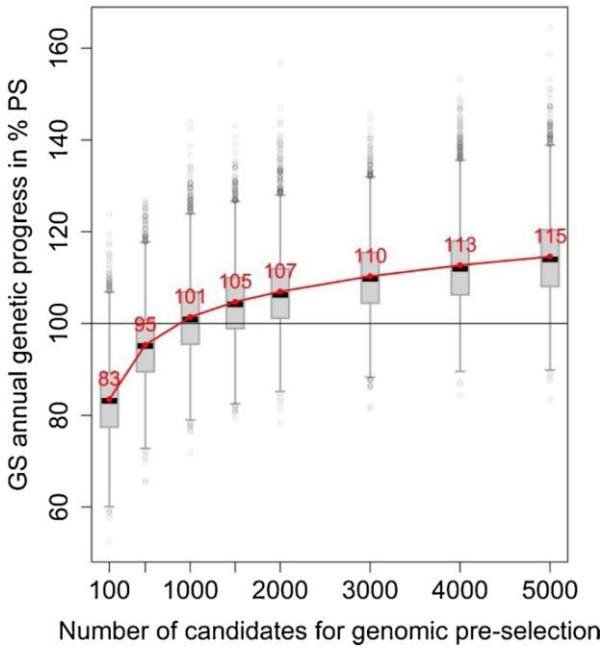


Fig. 7. Annual response to selection in the GS scheme, expressed in % of the annual selection response in the conventional PS scheme, according to the number of candidates subjected to genomic preselection. Values in red are means of 5000 replicates. The horizontal black line indicates annual selection response with GS equal to annual selection response with PS.

Figure 39 Progrès génétique sur la production de latex avec le schéma génomique suggéré pour l'hévéa (exprimé en pourcentage du progrès génétique avec le schéma phénotypique conventionnel), en fonction du nombre d'individus soumis à la sélection génomique (Cros et al., 2019)

2.4.2. Nouveaux schémas d'amélioration génétique

En me basant sur ces résultats et en m'appuyant sur les schémas actuellement utilisés par nos partenaires (c-à-d SRR impliquant deux groupes hétérotiques complexes pour le palmier à huile, et

sélection clonale au sein de familles biparentales pour l'hévéa), je suggère de nouveaux schémas d'amélioration pour le palmier à huile et l'hévéa.

Pour le **palmier à huile** (Figure 15, partie droite), dans le premier cycle, la SG est appliquée en pépinière sur une vaste population (étape B de la figure). Les prédictions sont obtenues avec un modèle calibré avec les données des cycles précédents et portent sur l'ensemble des caractères d'intérêt. Les individus les moins performants sont éliminés, et ceux restants constituent le champ semencier. Ceci représente une première amélioration par rapport au schéma traditionnel, dans lequel le champ semencier est composé de l'ensemble des individus disponibles. Les mêmes prédictions génomiques servent aussi à identifier les individus à tester en croisements hybrides (C), ce qui apporte une seconde amélioration au schéma, avec (i) un gain de temps permis par l'arrêt de l'évaluation des individus A et B sur leurs valeurs propres, (ii) une sélection avant essai sur un plus grand nombre de candidats, et (iii) une sélection avant essai sur un plus grand nombre de caractères. Les individus les plus performants en croisements hybrides sont ensuite sélectionnés pour produire une nouvelle génération (E, F). En parallèle, les meilleurs croisements hybrides entre ces individus sont réalisés pour produire une population soumise à la SG pour identifier les meilleures têtes de clones (G), dont l'évaluation finale est réalisée en essais clonaux (H). Ceci représente une troisième amélioration du schéma par rapport à sa version conventionnelle, avec (i) une sélection des têtes de clones à tester en essais clonaux plus rapide (plus de nécessité d'évaluation des valeurs propres), (ii) réalisée sur un plus grand nombre de candidats et (iii) sur un plus grand nombre de caractères.

Des variantes autour de ce schéma génomique restent possibles, et des études complémentaires permettraient de préciser davantage la meilleure stratégie. Par exemple, il est possible de conserver les étapes d'évaluation sur valeurs propres et d'utiliser ces informations dans les modèles de SG. Cela pourrait se faire comme dans Nyouma et al. (2020) pour la sélection des têtes de clones, ou, pour la sélection dans les groupes A et B, en s'appuyant sur les modèles développés chez les animaux pour combiner les données des parents et des hybrides (Christensen et al., 2014 ; Vitezica et al., 2016 ; Xiang et al., 2016). Ceci permettrait d'accroitre la précision de la SG (Nyouma et al., 2020), mais augmenterait le coût et la durée des cycles. Une autre variante consisterait à produire la population de candidats à la sélection (étape G) en parallèle aux essais hybrides, en se basant sur les meilleurs croisements prédits par le modèle génomique sans attendre l'analyse des essais hybrides. Ceci accélérerait encore la sélection clonale, mais avec une précision certainement moindre lors de la SG avant essais clonaux.

Dans les populations actuelles, le génotypage peut se faire avec 10 000 SNP, de préférence de puce, et est à réaliser sur les parents des croisements et un échantillon des individus hybrides (environ 1 000), répartis dans les différents croisements. La population d'application doit être représentative de la population de calibration (descendants, plein-frères). L'approche statistique actuellement recommandée pour les prédictions est le GBLUP ASGM (c-à-d sans prise en compte de l'origine parentale des allèles). L'optimisation de la population de calibration, par exemple avec le CDmean, permet d'améliorer la précision. La *mate selection*, en optimisant la sélection et le croisement entre individus sélectionnés, permet d'augmenter le gain génétique tout en limitant la hausse de la consanguinité dans les populations parentales. Le progrès génétique est aussi lié à la taille de la population d'application, qui dans l'idéal devrait comporter plusieurs milliers d'individus pour tirer un bénéfice significatif de la SG.

Pour l'**hévéa**, avec un schéma basé sur des familles de plein-frères considérées séparément, une population de calibration est nécessaire pour chaque famille. Le schéma génomique démarre donc par un essai clonal (SSCT) dont les données servent à calibrer un modèle de SG appliqué sur une population importante d'individus du même croisement, en pépinière (Figure 20, partie droite). Ceci remplace la sélection phénotypique actuellement réalisée à l'étape SET et améliore le schéma du fait que la SG est appliquée (i) sur un plus grand nombre d'individus, (ii) sur plus de caractères et (iii) avec une précision accrue par rapport à la sélection phénotypique en SET (voir 2.2.5.g). La nécessité d'avoir deux SSCT dans le premier cycle aboutit à ce que sa durée reste la même entre le schéma génomique et le schéma conventionnel. Cependant, au-delà du premier cycle, les cycles deviendront plus courts : un seul SSCT sera nécessaire, puisque le modèle de SG aura été calibré avec les données du premier cycle. En outre, la population de calibration utilisée dans le deuxième cycle comprendra les données agrégées des deux SSCT du premier cycle et, dans les cycles suivants, les données des nouveaux SSCT seront ajoutées à la population de calibration.

Dans un contexte de familles de plein-frères, les prédictions génomiques peuvent se faire avec 200 clones pour la calibration et 1 600 SNP de GBS.

Une alternative au type de schéma étudié serait d'utiliser une population plus complexe, et en particulier un ensemble de familles connectées par des parents communs. Les implications d'un tel changement restent à évaluer dans le contexte de l'hévéa. D'ici quelques années, une première étude sur cet aspect sera conduite dans le projet « IFC-CV4 » (voir 2.5). Ce type d'approche de SG avec des familles connectées est mis en œuvre dans un certain nombre d'espèces pérennes, notamment le pin *Pinus taeda*, l'épicéa, l'eucalyptus (Grattapaglia, 2017), le pommier (par exemple Kumar et al., 2015 ; Muranty et al., 2015) et les agrumes (Minamikawa et al., 2017). Cette approche a l'avantage de conduire à une population de calibration unique (et donc plus grande) au lieu de populations spécifiques à chaque famille. Cependant, cette augmentation de la taille de la population de calibration s'accompagne d'une diminution de l'apparentement avec la population d'application. Par conséquent, dans la pratique, une approche de SG utilisant une population complexe comprenant plusieurs familles risque d'être plus compliquée à gérer, d'autant que la biologie de l'hévéa (voir 2.2.5.g) pourrait causer des déséquilibres importants en termes de représentation des parents et de taille des familles, et que la précision de la SG varierait entre les candidats à la sélection en fonction de leur apparentement effectif avec les clones de calibration. Cela pourrait également conduire à une précision de SG inférieure à celle obtenue avec des populations de calibration spécifiques à chaque famille (Crossa et al., 2017 ; Lenz et al., 2017 ; Schopp et al., 2017 ; Toro et al., 2017 ; Würschum et al., 2017b).

2.5. Applications pratiques

Les résultats obtenus sur la SG du palmier à huile et de l'hévéa ont permis de proposer des **applications pilotes**. Elles ont été intégrées au projet Cirad-PalmElit « SelGen_Palm » que je porte et dans le projet « IFC-CV4 » porté par mon collègue Vincent Le Guen, et dans lequel j'interviendrai notamment sur les prédictions génomiques.

Pour l'hévéa, ce travail a démarré en 2021, avec la récolte d'un lot de 10 000 graines dans des plantations de l'IFC en Côte d'ivoire. Un génotypage sera réalisé sur les plantules afin de déterminer les parents mâles et de connaitre leur répartition entre familles. Deux ensembles de 300 individus seront évalués, chacun dans un essai clonal (CCGE). Ils serviront à calibrer le modèle de SG, pour une application dans le reste de la population, avec comme objectif la sélection de 100 clones. Les

génotypages seront réalisés par GBS. L'approche prévue ici représentant une innovation par rapport à ce qui a été étudié chez l'hévéa jusqu'à présent (c-à-d population complexe *versus* famille biparentale), des validations seront réalisées, notamment entre essais clonaux.

Pour le palmier à huile, il est prévu d'utiliser comme population de calibration les dispositifs expérimentaux sur lesquels j'ai travaillé jusqu'à présent. Une population d'application d'environ 1 000 individus est en cours de définition. Les génotypages seront réalisés par puce à SNP.

2.6. Conclusion

Il a été intéressant pour moi de travailler sur le palmier à huile et l'hévéa, car cela m'a permis de valoriser au mieux mes compétences, tout en me donnant la possibilité d'élargir mes sujets de réflexion. La combinaison de ces deux espèces est apparue pertinente car, au-delà du fait qu'elles sont souvent cultivées dans les mêmes zones et/ou par les mêmes sociétés de plantation, l'amélioration génétique de ces deux plantes pérennes présente plusieurs similitudes importantes : un schéma avec deux étapes d'évaluation (valeurs propres et performances en essai), des phases d'évaluation phénotypiques lourdes qui rendent les cycles de sélection longs et qui limitent le nombre d'individus évalués en essai (faible intensité de sélection), des évaluations en essai avec une précision élevée.

Des conclusions ressemblantes quant au potentiel de la SG ont été tirées. Les prédictions génomiques peuvent atteindre des précisions élevées mais, pour l'instant, la méthode n'est pas efficace pour tous les caractères d'intérêt, avec une faible précision (<0,2) et/ou sans être capable de capturer la ségrégation mendélienne. Sur ces deux espèces, les perspectives actuelles d'applications des prédictions génomiques sont l'utilisation de la SG à la place de la sélection phénotypique actuellement réalisée sur valeurs propres (en SET chez l'hévéa et dans le champ semencier chez le palmier à huile) avant les essais aux champs, que ce soit les tests sur descendance (prédiction de la valeur des croisements) chez le palmier à huile ou les essais clonaux (prédiction de la valeur génétique individuelle des candidats ortets) chez le palmier à huile et l'hévéa.

Pour les deux espèces, l'augmentation de l'efficacité du schéma découlera d'une plus grande intensité de sélection à l'échelle du cycle, avec l'application de la SG sur un plus grand nombre d'individus, sur plus de caractères et, au moins pour une partie des caractères, avec une précision accrue par rapport à la sélection phénotypique actuelle avant essais. Un léger raccourcissement du cycle est aussi possible chez le palmier à huile, ainsi qu'un raccourcissement plus important, mais à partir du second cycle, chez l'hévéa.

Des différences entre ces deux cas d'études existent aussi. Pour le palmier à huile, les populations d'amélioration sont complexes et structurées alors que chez l'hévéa, l'approche actuelle porte sur des familles de plein-frères. L'existence d'une filière de production de semences chez le palmier à huile fait que plus de moyens sont disponibles par rapport à l'hévéa, ce qui se concrétise par plus de ressources et d'outils pour la génomique (génomes de référence, puce à SNP). En contrepartie, les échanges scientifiques sont beaucoup plus faciles sur l'hévéa, comme en témoigne notre collaboration avec le Brésil et l'utilisation conjointe des mêmes jeux de données. Une autre différence résulte dans le type de sortie variétale, avec des croisements hybrides inter-populations chez le palmier à huile et des clones d'une population chez l'hévéa. L'amélioration clonale existe aussi chez le palmier à huile mais, pour l'instant, son importance est marginale. Ces différences ont justifié la conduite d'études sur des aspects spécifiques à l'une ou l'autre de ces deux espèces.

2.7. Liste des publications indexées

2.7.1. Articles

(en lien avec la sélection génomique)

Seyum E. G., Bille N. H., Abtew W. G., Munyengwa N., Bell J. M., **Cros D.**; *under review.* Genomic selection in tropical perennial crops and plantation trees: a review.

Seyum E.G., Bille H.N., Abtew W.G., Rastas P., …, **Cros D.**, *under review*. Genome properties of key oil palm (Elaeis guineensis Jacq.) breeding populations.

Tchounke B., Sanchez L., Bell J.M. et **Cros D**., *under review*. Mate selection: a useful approach to maximize genetic gain and control inbreeding in genomic and conventional oil palm (Elaeis guineensis Jacq.) hybrid breeding.

Nyouma A., Bell J.M., Jacob F., Riou V., Manez A., …, **Cros D.**, 2022. Improving the accuracy of genomic predictions in an outcrossing species with hybrid cultivars between heterozygote parents: case study of oil palm (Elaeis guineensis Jacq.). Mol. Genet. Genomics. https://doi.org/10.1007/s00438-022-01867-5 - **IF : 3.291**

Munyengwa N., Le Guen V., Bille H.N., Souza L.M., Clément-Demange A., …, **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. **- IF2020 : 5.736**

Nyouma A., Bell J.M., Jacob F., Riou V., Manez A., …, **Cros D.**, 2020. Genomic predictions improve clonal selection in oil palm (Elaeis guineensis Jacq.) hybrids. Plant Science, 299: 110547. - **IF2020 : 4.729**

Nyouma A., Bell J.M., Jacob F. et **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): 69. **- IF2020 : 2.297**

Cros D., Mbo-Nkoulou L., Bell J.M., Oum J., Masson A. et al., 2019. Within-family genomic selection in rubber tree (Hevea brasiliensis) increases genetic gain for rubber production. Industrial Crops and Products, 138: 111464. - **IF2020 : 5.645**

Cros D., Tchounke B. et Nkague-Nkamba L., 2018. Training genomic selection models across several breeding cycles increases genetic gain in oil palm in silico study. Molecular Breeding, 38(7): 89. **- IF2020 : 2.589**

Cros D., Bocs S., Riou V., Ortega-Abboud E., Tisné S. et al., 2017. Genomic preselection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses. BMC Genomics, 18(1): 839. - **IF2020 : 3.969**

Marchal A., Legarra A., Tisné S., Carasco-Lacombe C., Manez A., …, **Cros D**., 2016. Multivariate genomic model improves analysis of oil palm (Elaeis guineensis Jacq.) progeny tests. Molecular Breeding, 36(2): 1‑13. **- IF2020 : 2.589**

Bouvet J.M., Makouanzi G., **Cros D.**, Vigneron P. 2016. Modeling additive and non-additive effects in a hybrid population using genome-wide genotyping: Prediction accuracy implications. Heredity, 116 : p. 146-157.
https://doi.org/10.1038/hdy.2015.78. **- IF2020 : 3.821**

Cros D., Denis M., Bouvet J.-M. et Sanchez L., 2015. Long-term genomic selection for heterosis without dominance in multiplicative traits: case study of bunch production in oil palm. BMC Genomics, 16(1): 651. - **IF2020 : 3.969**

Cros D., Denis M., Sánchez L., Cochard B., Flori A. et al., 2015. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.). Theoretical and Applied Genetics, 128(3): 397‑410. - **IF2020 : 5.699**

Cros D., Sánchez L., Cochard B., Samper P., Denis M., Bouvet J.-M. et Fernández J., 2014. Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population. Theoretical and Applied Genetics, 127(4): 981-994. - **IF2020 : 5.699**

Carré C., Gamboa F., **Cros D.**, Hickey J.M., Gorjanc G., Manfredi E., 2013. Genetic prediction of complex traits: Integrating infinitesimal and marked genetic effects. *Genetica*, **141** (*4-6*) : p. 239-246. http://dx.doi.org/10.1007/s10709-013-9722-9 - **IF 2020 : 1.082**

2.7.2. Chapitres de livres

Soh A.C., Mayes S., Roberts J.A., **Cros D.** et Purba A.R., 2017. Breeding Plans and Selection Methods, pp. 143‑163 in Oil Palm Breeding: Genetics and Genomics, Soh A.C., Mayes S., Roberts J.A., Boca Raton, CRC Press. - **book chapter**

Soh A.C., Mayes S., Roberts J., …, **Cros D.,** 2017. Molecular genetics and breeding, pp. 225-281 in Oil Palm Breeding: Genetics and Genomics, Soh A.C., Mayes S., Roberts J.A., Boca Raton, CRC Press. - **book chapter**

2.7.3. Mémoires

Cros D. 2014. Etude des facteurs contrôlant l'efficacité de la sélection génomique chez le palmier à huile (*Elaeis guineensis* Jacq.). Montpellier : Montpellier SupAgro, 204 p.. Thèse de doctorat -- Systèmes intégrés en biologie, agronomie, géosciences, hydrosciences, environnement (SIBAGHE). Biologie intégrative des plantes (BIP), Thèse de doctorat -- Systèmes intégrés en biologie, agronomie, géosciences, hydrosciences, environn.

Partie 3. Projet de recherches et perspectives

3.1. Introduction

Mon projet de recherches pour les années à venir continuera de porter sur l'**optimisation des schémas d'amélioration génétique du palmier à huile et de l'hévéa par l'utilisation d'approches génomiques**. Ce choix s'appuie sur le fait que les résultats atteints jusqu'ici, chez ces deux espèces mais aussi chez les plantes et les animaux de façon générale, montrent que les schémas d'amélioration génomiques peuvent contribuer fortement à amener la production au niveau des besoins à venir, tout en minimisant les impacts environnementaux.

Je suis largement motivé par les possibles retombées pratiques de mon travail, mais j'ai aussi un intérêt pour une question plus théorique, liée au fait que les recherches sur la SG sont marquées par deux grands types d'approches conceptuellement différentes. Certaines approches tentent d'intégrer le plus possible les connaissances biologiques disponibles concernant la relation entre le génotype et le phénotype des caractères d'intérêt. Dès le départ, des modèles prédisant des effets associés aux marqueurs ont ainsi été développés (RRBLUP) avec de nombreuses variantes, destinées à intégrer la possibilité d'une distribution non-normale des effets des marqueurs (l' « alphabet Bayésien »). Par la suite, des modèles intégrant un déterminisme génétique non additif ont été développés (dominance, épistasie). Des méthodes pour tenir compte d'une grande diversité d'informations *a priori* sur les marqueurs ont aussi été mises au point : informations sur des QTL ou sur des gènes candidats associés aux marqueurs, informations d'annotations, etc. Aujourd'hui, une importance croissante est accordée aux prédictions multi-omiques, avec comme objectif de modéliser tous les niveaux fonctionnels liant le génotype au phénotype. Certaines méthodes actuelles de SG visant à tenir compte des interactions entre génotype et environnement suivent cette même logique, avec des prédictions intégrant les connaissances écophysiologiques existantes sur le caractère cible par l'intermédiaire de modèles de croissance (voir 3.5). Parallèlement à ces développements, de nombreuses approches de SG de types boîte noire, c-à-d sans considérer les mécanismes pouvant lier les génotypes aux phénotypes, ont aussi été testées : GBLUP, RKHS, *random forest*, *support vector machine*, ... (voir 2.2.3.c). Aujourd'hui, les réseaux de neurones artificiels (ANN) font l'objet d'un nombre croissant d'études de SG (voir 3.2). Il existe donc une évolution vers une sophistication de plus en plus grande au sein de ces deux grands types d'approches, qui les font diverger, avec de plus en plus de connaissances biologiques dans un cas et des modèles boîtes noires de plus en plus complexes dans l'autre. Je suis particulièrement intéressé par les futures recherches qui montreront laquelle de ces deux approches est la plus efficace pour ce qui est d'optimiser en pratique le rythme du progrès génétique, et je souhaite par mon projet de recherches apporter une contribution dans ce domaine. Evidemment, il est aussi possible que l'avenir soit à la fusion entre ces deux approches, à l'image des ANN multi-couches modélisant les différents niveaux -omiques (3.2).

Pour mettre en œuvre mon projet de recherches, je continuerai à m'investir dans des activités d'**encadrement scientifique**, en prenant plus de responsabilités, notamment avec la direction de doctorats et l'encadrement de post-docs. Dans un premier temps, j'accompagnerai Billy Tchounke et Essubalew Seyum jusqu'à leur soutenance, prévue pour fin 2023 et fin 2022, respectivement. Billy Tchounke travaille actuellement sur des simulations avec lesquelles il étudie comment le progrès génétique sur le poids de régimes est affecté par l'utilisation d'indices de sélection combinant le nombre et le poids moyen des régimes. A partir du dernier trimestre 2022, je co-encadrerai un nouveau doctorant, Daouda Kouassi, avec le Professeur Akaffou, de l'université de Daloa (Côte

d'ivoire). Il travaillera sur l'optimisation de la SG appliquée à l'hévéa. Je me suis impliqué dans la construction de son sujet, en concertation avec mes collègues généticiens de l'hévéa, Vincent Le Guen et André Clément Demange, et Daouda, qui est actuellement responsable des expérimentations agronomiques dans une grande société de plantation. Le travail de Daouda fera suite aux premières études que nous avons conduites sur l'hévéa. Il étudiera de nouveaux aspects, notamment grâce à des données sur un essai évaluant des clones choisis sur des prédictions génomiques au sein des descendants de la famille utilisée jusqu'à présent, et grâce à des données sur une nouvelle famille.

Concernant les **nouvelles espèces** sur lesquelles je pourrai m'investir, je m'associerai notamment à des collègues spécialistes de ces cultures, pour apporter un soutien spécifique sur les aspects de prédiction génomique. Je me suis ainsi déjà impliqué dans la rédaction d'un *workpackage* sur la SG du teck (*Tectona grandis*) dans un projet porté par Jean Marc Gion (UMR AGAP) et soumis récemment. Je collabore aussi avec le Professeur Achigan-Dako (UAC) pour le co-encadrement scientifique de Luther Nkoulou dans son doctorat sur la SG du bananier. A cette occasion, j'ai collaboré avec Guillaume Martin (UMR AGAP), qui a développé le logiciel VcfHunter qui permet de faire l'appel des SNP du GBS chez les polyploïdes.

Je continuerai de mener mes recherches en combinant analyses sur **données expérimentales** et par **simulations informatiques**. Je n'utiliserai par contre plus HaploSim mais un outil de simulation plus performant. HaploSim était l'un des meilleurs outils au moment de mon doctorat, mais il souffre de plusieurs limites : simulation d'un déterminisme génétique purement additif, besoin important en mémoire, relative lenteur en considérant des populations de taille importante avec beaucoup de marqueurs, et peu de fonctions prédéfinies. Aujourd'hui, de nouveaux outils ont été développés. En 2021, j'ai mis en place un groupe de travail sur la simulation de schémas d'amélioration (voir 1.5.2) avec comme objectif d'identifier un outil plus satisfaisant qu'HaploSim. Après une étape de recensement des besoins des membres du groupe et des outils disponibles, nous sommes actuellement en train de tester MoBPS (Pook et al., 2020) et AlphaSimR (Gaynor et al., 2021).

J'ai déjà mentionné dans la synthèse de mes travaux de recherche plusieurs **aspects d'intérêt que je considérerai dans le prolongement des recherches que j'ai mené jusqu'ici**. Il s'agit notamment, pour le palmier à huile, de l'utilisation de modèles combinant les données des parents et des hybrides (Christensen et al., 2014 ; Vitezica et al., 2016 ; Xiang et al., 2016), et de la modélisation des effets de dominance développée par González-Diéguez et al. (2021). Aussi, chez le palmier à huile et l'hévéa, les différentes variantes possibles des schémas d'amélioration génomiques suggérés mériteraient d'être étudiées plus en détail (voir 2.4.2). Il serait intéressant d'évaluer les performances de modèles multi-variés sur d'autres ensembles de caractères que ceux considérés jusqu'à présent. Le progrès génétique par unité de coût des différents schémas doit aussi être comparé.

En dehors de ces aspects directement liés à mon travail actuel, je compte aussi m'intéresser à de **nouveaux aspects**, que j'ai identifié parmi les nombreuses recherches conduites sur l'amélioration de la méthodologie de la SG, tant chez les plantes que chez les animaux, et notamment sur les approches statistiques de prédiction et sur les données à fournir aux modèles. Dans la suite de ce document, je présenterai parmi ces aspects ceux qui me semblent les plus prometteurs et que je prévois d'explorer. Ils apparaissent par ordre de faisabilité. Les données actuellement disponibles sur le palmier à huile et l'hévéa permettent déjà de travailler sur les points 3.2, 3.3, 3.4.1 et 3.4.2 ; ainsi que sur une partie du point 3.4.3. Le travail a d'ailleurs déjà démarré sur certains aspects. Les points 3.4.4 à 3.6 nécessitent par contre l'acquisition de données spécifiques.

Pour mener à bien ce projet de recherches, je continuerai de m'appuyer sur des **projets** que je porterai, avec PalmElit et ses partenaires pour le palmier à huile, et portés par mes collègues de

l'hévéa; avec l'IFC. Je souhaite aussi élargir les projets que je porte. J'essaierai notamment de monter des projets multi-espèces, sur des aspects méthodologiques transversaux, qui devraient permettre d'accéder à de nouvelles sources de financements ; à l'image du projet SelGen_3D impliquant palmier à huile, hévéa, cacaoyer et eucalyptus, pour lequel j'ai obtenu un financement du Cirad (voir 2.1.3 et 3.4.2).

Enfin, je continuerai d'**accompagner nos partenaires privés dans l'application pratique de la SG** au sein de leur schéma d'amélioration.

3.2. Réseaux de neurones artificiels

Les méthodes de *machine learning* sont des approches complexes de type boîte noire qui font l'objet d'un intérêt croissant pour les prédictions génomiques, car elles ont plusieurs caractéristiques particulièrement intéressantes dans ce contexte. Elles évitent l'utilisation d'hypothèses qui souvent ne sont pas respectées (Gianola et van Kaam, 2008), et elles sont bien adaptées pour tenir compte des effets non additifs et pour les prédictions multi-omiques, c-à-d utilisant des données provenant de différentes sources biologiques (voir 3.4.3) (Bayer et al., 2021 ; Montesinos-López et al., 2021 ; Tong et Nikoloski, 2021). Le RKHS est l'approche de *machine learning* la plus souvent évaluée pour la SG. Chez les plantes pérennes tropicales, des exemples sont notamment disponibles pour le bananier (Nyine et al., 2018), l'eucalyptus (Rambolarimanana et al., 2018 ; Tan et al., 2017) et l'hévéa (Cros et al., 2019). Plusieurs autres méthodes de *machine learning* ont été mises en œuvre pour les prédictions génomiques, comme le *support vector machine*, la *random forest* et les réseaux de neurones artificiels (ANN) (Figure 40) (Long et al., 2011 ; Montesinos López et al., 2022 ; Montesinos-López et al., 2021 ; Tong et Nikoloski, 2021). Des résultats mitigés ont été obtenus jusqu'à présent et ont abouti à ce que le GBLUP soit la méthode la plus largement utilisée actuellement. Une des raisons pourrait être que les populations d'entraînement n'étaient pas suffisamment grandes pour permettre un entraînement optimal de ce type d'approche (Montesinos-López et al., 2021).

Cependant, le développement des capacités d'acquisition de données multi-omiques, la construction de populations de calibration de plus en plus grandes, et des résultats positifs obtenus récemment avec les ANN justifient que le potentiel de ces méthodes soient étudiées plus en détail. Maldonado et al. (2020) ont comparé plusieurs modèles classiques de prédiction paramétriques avec le RKHS et deux approches d'ANN, dont le *long short-term memory network* (LSTM). Le LSTM est un type d'ANN récurrent avec un grand nombre de couches de neurones (*deep learning*). Sur des jeux de données d'*Eucalyptus globulus* et du maïs, ils ont montré que les prédictions faites avec la méthode LSTM étaient significativement plus précises pour tous les caractères considérés (voir Figure 41 montrant leurs résultats chez le maïs), à partir du moment où une optimisation appropriée des hyperparamètres de la méthode était réalisée. Chez le café, Sousa et al. (2022 ; 2021) ont montré que des approches ANN donnaient de meilleures précisions que le GBLUP pour les différents caractères étudiés, c-à-d la résistance à la rouille (*C. arabica* et *C. canephora*) et le rendement (*C. canephora*).

L'évaluation de ce type de méthode semble intéressante pour le palmier à huile et l'hévéa et peut se faire avec les jeux de données disponibles, les populations de calibration de Maldonado et al. (2020) et Sousa et al. (2022 ; 2021) ayant des tailles du même ordre de grandeur. La mise en œuvre de ce volet demandera de ma part un effort de formation sur le *deep learning,* ce type d'approche ne m'étant pas familier. Il existe cependant aujourd'hui des packages qui permettent d'appliquer du *deep learning* dans un contexte de prédiction génomique, tel Keras et DeepGP (Montesinos-López et al.,

2021 ; Pérez-Enciso et Zingaretti, 2019 ; Zingaretti et al., 2020), sur lesquels je pourrai m'appuyer. Je chercherai aussi à m'associer à d'autres chercheurs avec des compétences dans ce domaine. Par ailleurs, j'ai inclus la possibilité d'une demande de poste d'un chercheur ayant des compétences sur les ANN dans la réflexion menée au sein de l'équipe GSP sur la définition de nos besoins de recrutements. Dans 5 ou 10 ans, la conduite de mon projet de recherche dira si mon investissement sur cette méthodologie, actuellement très populaire, m'aura permis de faire progresser l'intérêt pratique de la SG ou aura fait de moi une *fashion victim*. Quoiqu'il en soit, un investissement sur le *deep learning* ne pourra pas être vain, car les compétences développées dans ce domaine pourront être mises à profit utilement, d'une façon ou d'une autre, car le *deep learning* a un très vaste champ d'application.

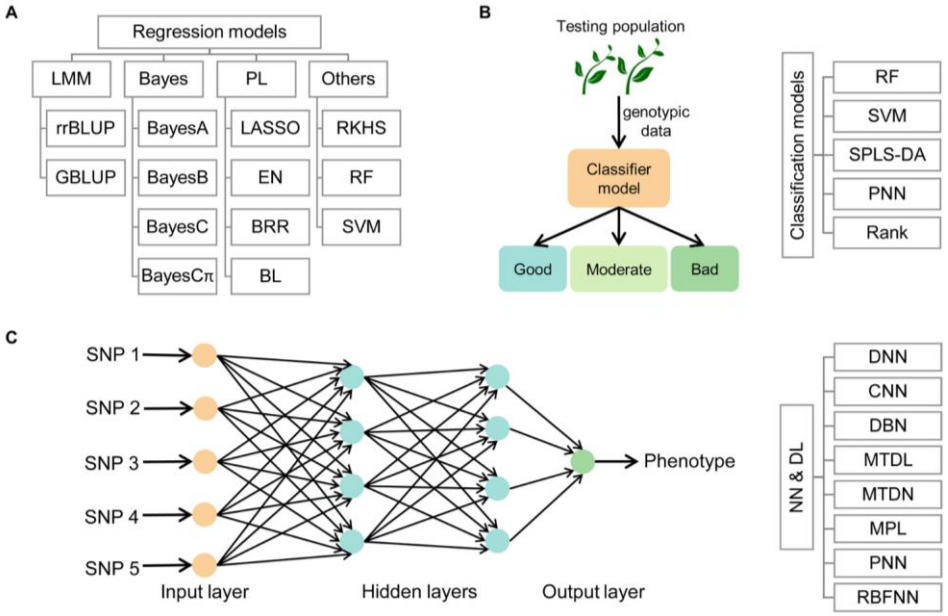


Fig. 2. Overview of statistical approaches for genomic selection. The genomic selection models can be classified into three groups based on statistical techniques used, namely (A) regression, (B) classification, and (C) deep learning. LMM: linear mixed model; rrBLUP: ridge regression best linear unbiased prediction; GBLUP: genomic best linear unbiased prediction; PL: penalized linear model; LASSO: least absolute shrinkage and selection operator; EN: elastic net; BRR: Bayesian ridge regression; BL: Bayesian LASSO; RKHS: reproducing kernels Hilbert spaces regression; RF: random forest; SVM: support vector regression; SPLS-DA: sparse partial least squares discriminant analysis; NN: neural network; DL: deep learning; DNN: deep neural network; CNN: convolutional neural network; DBN: deep belief network; MTDL: multi-trait deep learning; MTDN: multivariate poisson deep learning; MPL: multilayer perceptron; PNN: probabilistic neural network; RBFNN: radial basis function neural networks.

Figure 40 Méthodes de prédictions génomiques par type d'approche statistique. A : régression, B : classification, C : réseaux de neurones artificiels (Tong et Nikoloski, 2021)

TABLE 3 | Estimates of predictive ability of complex traits for different genomic models assessed in 6 years old eucalypt trees.

Model/traits	WD	ST	BQ	TH	DBH
Bayes A	0.267^{e}	0.376^{d}	0.216^{d}	0.304^{c}	0.352^{e}
Bayes B	0.295^{d}	0.518^{b}	0.128^{g}	0.319^{c}	0.341^{e}
Bayes Cπ	0.455^{b}	0.544^{ab}	0.162^{f}	0.441^{b}	0.394^{d}
BL	0.301^{cd}	0.200^{e}	0.056^{h}	0.204^{d}	0.169^{g}
BRR	0.321^{c}	0.481^{c}	0.309^{c}	0.303^{c}	0.444^{c}
GBLUP	0.187^{g}	0.226^{e}	0.142^{g}	0.159^{e}	0.220^{f}
RKHS	0.223^{f}	0.225^{e}	0.180^{e}	0.197^{d}	0.230^{f}
BRNN	0.454^{b}	0.469^{c}	0.419^{b}	0.491^{a}	0.490^{b}
DL	0.471^{a}	0.557^{a}	0.460^{a}	0.496^{a}	0.556^{a}
$\hat{h}^2$ (SE)*	0.09 (0.05)	0.01 (0.03)	0.05 (0.04)	0.04 (0.04)	0.01 (0.03)

The study traits were pilodyn penetration (WD), stem straightness (ST), branch quality (BQ), tree height (TH), and diameter at breast height (DBH) of eucalypt trees. Predictive ability values followed by a common letter are not significantly different according to the Tukey–Kramer test at a level of significance of 0.01.
**Narrow-sense heritability. SE, standard error.*

Figure 41 Précision de prédiction en fonction du modèle de SG et du caractère chez le maïs (Maldonado et al., 2020)
DL : deep learning (approche LSTM)

3.3. Recombinaisons ciblées à partir des profils d'effets aux marqueurs

Comme déjà mentionné, un des facteurs limitant la sélection des cultures pérennes est la taille restreinte de la population de candidats à la sélection, car plus la population est grande, plus la recherche d'individus élites au sein de la diversité générée par la méiose est exhaustive. La SG permet d'avoir une population de candidats à la sélection plus grande en remplaçant le phénotypage par le génotypage. Le contrôle des gamètes générés lors de la méiose a le potentiel d'augmenter encore l'efficacité du schéma de sélection. Cela pourrait être possible en combinant les profils des effets des marqueurs obtenus à l'échelle du génome par des modèles de SG, et la recombinaison ciblée (Bernardo, 2017). Les profils des effets des marqueurs le long des chromosomes d'individus hétérozygotes peuvent en effet être utilisés pour identifier des sites où les recombinaisons maximiseraient la valeur génétique des gamètes en agrégeant des blocs d'allèles favorables (Figure 42). Les recombinaisons pourraient être obtenues sur ces sites par édition du génome, puis les individus édités régénérés seraient croisés pour produire une descendance qui devrait permettre d'obtenir un progrès génétique accru (Bernardo, 2017 ; Brandariz et Bernardo, 2019). Des outils d'édition du génome sont en cours de développement actif chez diverses plantes pérennes tropicales, notamment avec CRISPR/Cas9 chez le palmier à huile (Yarra et al., 2020 ; Yeap et al., 2021) et l'hévéa (Dai et al., 2021 ; Fan et al., 2020). Cependant, des études sont nécessaires chez ces espèces pour développer des approches efficaces de recombinaison ciblée et pour évaluer l'efficacité relative des schémas de sélection impliquant des recombinaisons ciblées et des schémas conventionnels.

Je prévois d'étudier l'effet de cette stratégie par simulation informatique. J'ai d'ailleurs inscrit ce travail pour le teck dans le projet Bioteak. Dans le cadre du groupe de travail sur la simulation des schémas d'amélioration, j'ai récemment démarré la simulation d'un schéma d'amélioration clonale en deux étapes (sélection sur valeurs propres individuelles puis sur valeurs clonales), correspondant à ce qui se fait chez l'hévéa et le teck. Ce script servira de base à étudier, entre autres choses, cet aspect.

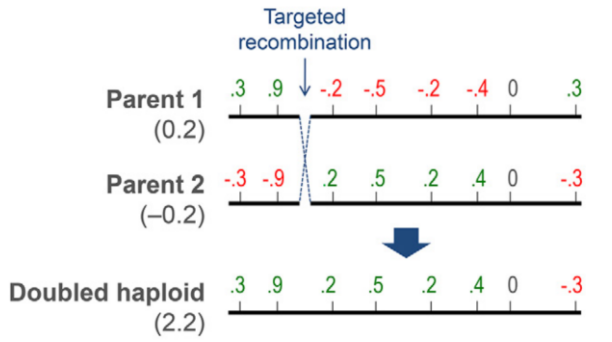


Fig. 1. Targeted recombination on the basis of genome-wide marker effects. The genotypic values of two parental inbreds and a doubled haploid developed via targeted recombination are in parentheses.

Figure 42 Utilisation des recombinaisons ciblées à partir des profils d'effets aux marqueurs (Bernardo, 2017)

3.4. Au-delà des données moléculaires par marqueur

3.4.1. Haploblocs

L'utilisation d'haploblocs composés de deux SNP adjacents ou plus, au lieu de SNP uniques, a été étudiée pour les prédictions génomiques, car elle pourrait augmenter la précision en capturant mieux l'identité par descendance entre les individus, en donnant un DL plus élevé entre les QTL et les allèles de l'haplobloc, ou en capturant les effets épistatiques entre les SNP dans le même haplobloc (Bhat et al., 2021 ; Goddard et Hayes, 2007 ; Hess et al., 2017). Par exemple, Ballesta et al. (2019) ont

exploré les avantages de l'utilisation de données haplotypiques pour la SG chez *Eucalyptus globulus* et ont montré que la précision de la prédiction était significativement plus élevée pour les caractères à faible héritabilité lorsque des haploblocs étaient utilisés au lieu de SNP uniques. Cependant, l'efficacité relative de l'utilisation d'haploblocs ou de SNP uniques pour les prédictions génomiques est affectée par de nombreux paramètres, en particulier la taille de la population de calibration, le niveau de DL, la méthode utilisée pour définir les haploblocs et la précision du phasage (Bhat et al., 2021 ; Goddard et Hayes, 2007 ; Hess et al., 2017).

Une comparaison de la précision de modèles de SG utilisant des haploblocs et des SNP seuls est possible avec les données que nous avons sur le palmier à huile et l'hévéa. L'utilisation d'haploblocs requiert le phasage des données moléculaires. Je réalise cette étape avec le logiciel Beagle (Browning et al., 2021), très largement utilisé compte tenu de ses performances en termes de précision, de vitesse de calcul et de capacité à gérer de gros jeux de données. L'implémentation d'un modèle GBLUP avec des haplotypes définis sur des données moléculaires obtenues avec un puce à SNP a été initiée par Achille Nyouma sur le palmier à huile, dans le cadre d'un CDD qu'il a effectué avec moi suite à son doctorat (voir 1.8.1). Il a utilisé deux méthodes pour la définition des haploblocs : les fenêtres distinctes (*distinct windows*) et le DL (Figure 43) (Ferdosi et al., 2016 ; Hickey et al., 2013 ; Teissier et al., 2020). La méthode des fenêtres distinctes consiste à construire des haploblocs en agrégeant un nombre fixe de SNP adjacents le long du chromosome. Dans la méthode basée sur le DL, un haplobloc est défini comme un groupe de SNP dans lequel le DL entre chaque paire de SNP est supérieur ou égal à un seuil fixé. Ce travail est en cours.

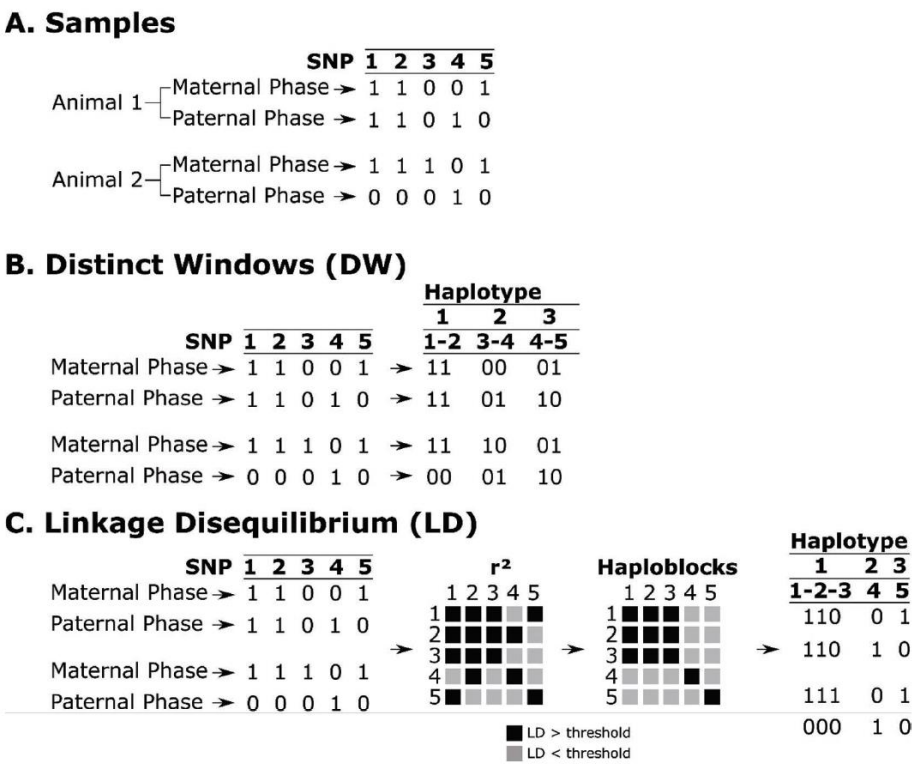


Figure 1. Construction of haplotypes using the distinct windows (DW) or linkage disequilibrium (LD) methods. Initially, genotypes are phased (A). In DW (B), the size of the window required to create the haplotypes needs to be defined (here, 2 SNP). In LD (C), LD between SNP needs to be estimated before construction of the haplotypes.

Figure 43 Illustration de la définition des haplotypes selon les méthodes « *distinct windows* » et « DL » (Teissier et al., 2020)

3.4.2. Information a priori sur les marqueurs

Une autre façon d'améliorer la précision de la SG est d'incorporer dans le modèle de prédiction les informations existantes concernant les polymorphismes, notamment celles obtenues à partir d'études de détection de QTL (Xu et al., 2020). Différentes approches de modélisation ont été développées à cette fin, et leur efficacité a été démontrée dans des études sur les animaux et les plantes (voir par exemple chez l'abricotier dans Nsibi et al. (2020)). Cependant, très peu d'études se sont penchées sur cet aspect chez les plantes pérennes tropicales jusqu'à présent. Chez le palmier à huile, Kwong et al. (2017a) ont appliqué le RRBLUP en utilisant uniquement les SNP ayant le score d'association GWAS le plus élevé, ce qui a permis de réduire la densité de marqueurs tout en obtenant une précision meilleure ou identique à celle obtenue en utilisant tous les SNP. Cependant, ces approches dépendent fortement de la définition des populations de calibration et d'application. Ainsi, dans le cacao, l'inclusion des SNP détectés par GWAS comme effets fixes dans le modèle de SG n'a pas amélioré les précisions de prédiction, ce qui résulte probablement d'une différenciation génétique trop élevée entre les populations d'entraînement et d'application, rendant les SNP détectés non pertinents (McElroy et al., 2018). Cet aspect mérite d'être étudié plus en détail sur le palmier à huile et l'hévéa.

Un grand nombre d'informations utilisables dans les modèles de SG sont disponibles chez ces deux espèces. De nombreux QTL ont ainsi été publiés : sur la production de latex, la croissance végétative et la résistance aux maladies chez l'hévéa (voir par exemple Conson et al., 2018 ; Rosa et al., 2018 ; Tran et al., 2016), et sur les composantes du rendement en huile, la croissance végétative et la résistance aux maladies chez le palmier à huile (voir par exemple Babu et al., 2021 ; Billotte et al., 2010 ; Daval et al., 2021 ; Pootakham et al., 2015 ; Teh et al., 2020 ; Tisné et al., 2015 ; Tisné et al., 2017). Lors du stage de M2 de Jean Oum, nous avons d'ailleurs évalué trois méthodes intégrant des informations sur des QTL de production de latex détectés au préalable par mes collègues de l'hévéa dans la même famille : le W-BLUP (Zhao et al., 2014) et le MultiBLUP (Speed et Balding, 2014), des extensions du RRBLUP et du GBLUP, respectivement, permettant de regrouper les marqueurs en classes (correspondant dans notre cas aux SSR dans les QTL et aux SSR hors QTL) ; et la modélisation en effets fixes des SNP situés dans les QTL (Bernardo, 2014). Ceci n'a pas amélioré la précision de la SG par rapport au RRBLUP, mais ce résultat a pu être la conséquence des faibles valeurs de R^2 associées aux QTL (9% - 19%), d'un nombre d'individus utilisés pour la détection de QTL trop restreint (<200) et de la faible densité de marquage (<350 SSR). Ce travail sera repris, notamment avec les données de génotypage GBS.

D'autres informations que celles fournies par des études de détection de QTL peuvent aussi être intégrées aux modèles de SG. Ainsi, pour les approches utilisant des classes de marqueurs (Sørensen et al., 2014 ; Speed et Balding, 2014 ; Zhao et al., 2014), les annotations du génome peuvent être utilisées pour définir les classes (marqueurs dans les introns, les exons, intergéniques ; marqueurs dans les gènes d'une voie de biosynthèse particulière, etc. ; voir Figure 44). Des informations d'annotation sont disponibles chez le palmier à huile et l'hévéa, notamment sur des gènes en lien avec la biosynthèse des triglycérides chez le palmier à huile (Amiruddin et al., 2020 ; Chan et al., 2017 ; Zheng et al., 2019) et du caoutchouc chez l'hévéa (Bini et al., 2022 ; Long et al., 2021 ; Nakano et al., 2021 ; Yamashita et Takahashi, 2020).

En dehors de ce type d'information que j'intégrerai dans des prédictions génomiques, je pense qu'il pourrait aussi être intéressant de tenir compte de la structure 3D du génome. Au sein du noyau, l'ADN présente une organisation 3D multi-échelle complexe définissant des territoires chromosomiques, des compartiments et, à une échelle <1Mbp, des *topologically associating domains*

(TAD), qui jouent un rôle clé dans la régulation des gènes (Long et al., 2020 ; Mota-Gómez et Lupiáñez, 2019 ; Ouyang et al., 2020 ; Szabo et al., 2019 ; Wang et al., 2020). Les TAD semblent rassembler des gènes impliqués dans les mêmes processus biologiques et avec des profils d'expression similaires, leur permettant d'être régulés conjointement, et notamment de passer ensemble de l'état inactif à l'état actif ou inversement (Long et al., 2020). Afin d'étudier cet aspect, je porte le projet SelGen_3D, qui vise à acquérir le profil de TAD sur plusieurs espèces de mon équipe (palmier à huile, hévéa, cacaoyer, eucalyptus) et à étudier l'intérêt de l'utilisation de ce type d'informations en prédictions génomiques.

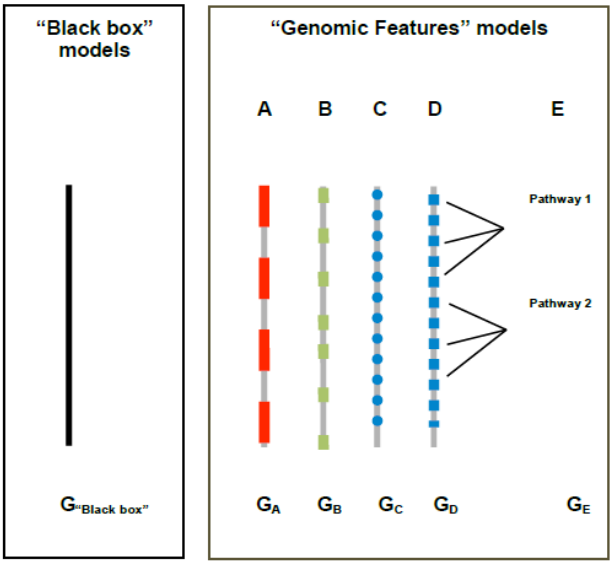


Figure 1. The "Black box" modeling approach works on the individual SNP level and treats all SNPs equally. The "Genomic feature" modeling approach accounts for the correlations among SNPs by grouping them according to genomic features such as A) location in transcriptionally active genomic regions, B) sequence-based prediction of deleteriousness, C) location in genomic regions found to be associated to the complex trait in previous GWAS studies, D) location in coding (exons, introns), non-coding or regulatory (e.g. promoters, transcription factor binding sites) regions, or E) location in genes part of biological complexes (PPI), pathways (KEGG) or modules (co-expression). Genomic parameters (G) such as variances, correlations and heritabilities are estimated for each layer of information using statistical models.

Figure 44 Principe du *genomic feature model* (Sørensen et al., 2014)

3.4.3. Endophénotypes

L'incorporation d'endophénotypes, c-à-d de phénotypes intermédiaires, dans les modèles de prédiction est une autre voie prometteuse des recherches sur la SG. Les endophénotypes, et en particulier les données transcriptomiques et métabolomiques, ont été utilisés conjointement avec les données génomiques chez plusieurs espèces végétales (Scossa et al., 2021 ; Tong et Nikoloski, 2021 ; Xu et al., 2020). Ces approches de prédictions multi-omiques devraient permettre de mieux saisir les effets très faibles et les effets non additifs, et de mieux modéliser la relation entre les génotypes et les phénotypes. Elles ont en particulier donné de bons résultats chez les hybrides, comme chez le maïs, le

riz et le colza, où elles ont surpassé les prédictions basées uniquement sur les seules données génomiques (Knoch et al., 2021 ; Scossa et al., 2021 ; Xu et al., 2020).

Dans mes premières simulations (Cros et al., 2015a) puis avec des étudiants de M2 (Alexandre Marchal et Evrard Akpla), nous avons mis en œuvre des analyses bi-variées avec un modèle correspondant à la juxtaposition de modèles univariés qui se retrouvent associés par des matrices de variance-covariance entre caractères pour les effets génétiques et résiduels (Mrode, 2014, p. 70). Une telle approche peut être mise en œuvre pour tenir compte d'endophénotypes. Elle peut cependant causer des difficultés en termes de calculs lorsqu'elle est appliquée sur de grandes populations et/ou de nombreux caractères, avec en particulier des problèmes de convergence et une forte augmentation du temps de calcul. Michel et al. (2018) ont utilisé une approche alternative qui permet de prendre en compte des endophénotypes avec un modèle univarié et un indice de sélection, tout en aboutissant à des précisions de prédiction plus élevées qu'avec un modèle multi-varié. Cette approche passe par deux étapes. Dans un premier temps des modèles univariés classiques (Équation 1) sont appliqués pour obtenir une prédiction génomique des valeurs génétiques des individus des populations de calibration et de validation pour les endophénotypes. Ces valeurs sont ensuite incluses comme effet fixe dans un modèle du même type, appliqué cette fois au caractère d'intérêt. Les valeurs génétiques prédites pour ce caractère d'intérêt valent finalement, pour un individu i : $\sum_j x_{ij} b_j + g_i$, avec g_il'effet génétique aléatoire de i pour le caractère d'intérêt, x_{ij} la valeur génétique prédite dans la première étape pour l'endophénotype j et b_j l'effet fixe estimé pour l'endophénotype j. Cette approche a augmenté les précisions de SG dans différentes études, notamment sur le blé (Michel et al., 2018) et l'abricot (Nsibi et al., 2020). Elle serait particulièrement intéressante à tester sur l'hévéa, pour lequel nous avons des données de production de latex mais aussi de saccharose, un précurseur indispensable de la synthèse du caoutchouc.

Pour le palmier à huile, il serait possible d'appliquer la méthode développée par Campbell et al. (2021) à la prédiction de la teneur en huile dans la pulpe des fruits. Leur approche est adaptée à un caractère d'intérêt qui peut se décomposer en une somme d'endophénotypes. La prédiction génomique du caractère d'intérêt est obtenue avec un modèle univarié où les différents endophénotypes, représentés chacun par un effet aléatoire associé à une matrice spécifique d'apparentements génomiques, sont additionnés (*multi-kernel model*). Les matrices d'apparentements sont calculées en affectant un poids à chaque SNP, les poids correspondants aux effets des marqueurs obtenus par des modèles de SG appliqués initialement aux différents endophénotypes. Campbell et al. (2021) ont mis en œuvre cette approche chez l'avoine pour prédire la teneur en lipides totaux des graines, en utilisant comme endophénotypes la teneur des différents acides gras. Ils ont obtenu des précisions de SG dépassant de plus de 10% les précisions obtenues avec un modèle multi-varié considérant conjointement la teneur en lipides totaux et la teneur des différents acides gras (Figure 45), et un modèle univarié incluant uniquement la teneur en lipides totaux. Cette approche peut être directement appliquée à la teneur en huile de la pulpe des fruits du palmier à huile. Il faut par contre acquérir de nouvelles données, c-à-d des profils individuels d'acides gras au sein de la population de calibration, afin de réaliser une première comparaison avec les approches actuelles.

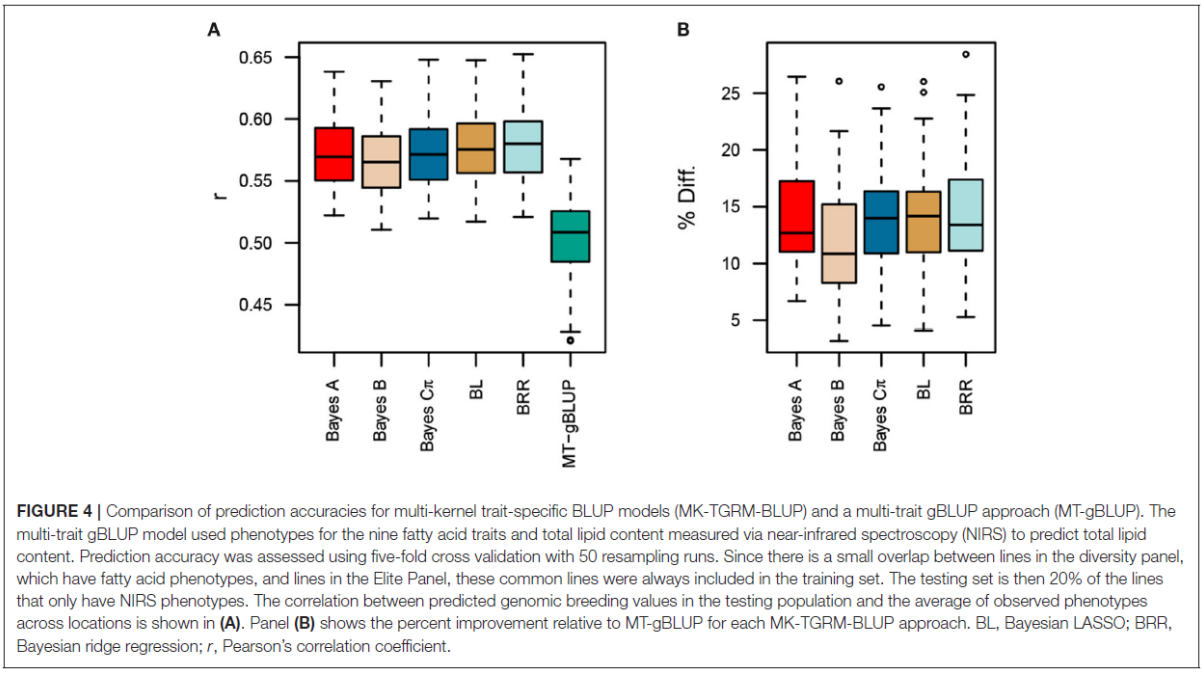
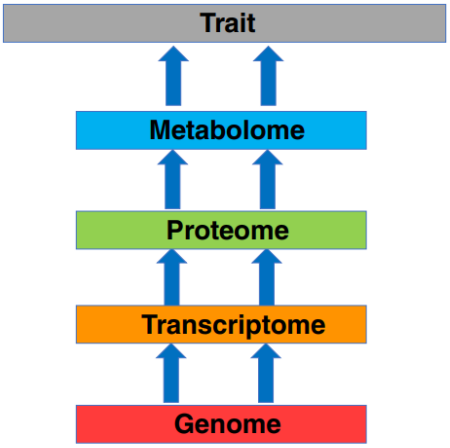


FIGURE 4 | Comparison of prediction accuracies for multi-kernel trait-specific BLUP models (MK-TGRM-BLUP) and a multi-trait gBLUP approach (MT-gBLUP). The multi-trait gBLUP model used phenotypes for the nine fatty acid traits and total lipid content measured via near-infrared spectroscopy (NIRS) to predict total lipid content. Prediction accuracy was assessed using five-fold cross validation with 50 resampling runs. Since there is a small overlap between lines in the diversity panel, which have fatty acid phenotypes, and lines in the Elite Panel, these common lines were always included in the training set. The testing set is then 20% of the lines that only have NIRS phenotypes. The correlation between predicted genomic breeding values in the testing population and the average of observed phenotypes across locations is shown in **(A)**. Panel **(B)** shows the percent improvement relative to MT-gBLUP for each MK-TGRM-BLUP approach. BL, Bayesian LASSO; BRR, Bayesian ridge regression; *r*, Pearson's correlation coefficient.

Figure 45 Précision des prédictions génomiques avec des modèles *multi-kernel* (BayesA à BRR) et un modèle multi-varié (MT-gBLUP) (Campbell et al., 2021)
Les différents modèles *multi-kernel* renvoient à la méthode d'estimation des effets aux marqueurs.

De façon générale, la méthode la plus couramment utilisée pour réaliser des prédictions multi-omiques repose sur l'utilisation de modèles *multi-kernel*. Ce type de modèle comprend plusieurs effets aléatoires qui sont chacun associé à une matrice de variance-covariance proportionnelle à une matrice d'apparentement, dont les éléments sont calculés selon une méthodologie (*kernel*) spécifique à chaque terme (Figure 46) (Rice et Lipka, 2021, p. 5). Christensen et al. (2021) ont développé une approche alternative, le *Genomic Omics BLUP* (Christensen et al., 2021), qui considère conjointement les différents niveaux -omiques. Les réseaux de neurones artificiels (ANN) peuvent aussi être mis à profit, car ils sont particulièrement bien adaptés pour incorporer dans un même modèle un grand nombre de données multi-omiques (Montesinos-López et al., 2021). Par exemple, Zhao et al. (2022) a développé pour la prédiction multi-omique une approche baptisée *neural networks linear mixed model* (NN-LMM), reposant sur un réseau de neurone à couches multiples dans lequel les génotypes affectent les niveaux -omiques intermédiaires, qui eux-mêmes régulent les phénotypes (Figure 47). Cette approche utilise les modèles linéaires mixtes habituels de la SG (GBLUP, BayesA, BayesB, etc.) pour estimer les effets des marqueurs ou des éléments des autres niveaux -omiques, tient compte des relations non-linéaires existant entre les niveaux -omiques intermédiaires et les phénotypes, et gère l'hétérogénéité des données manquantes entre niveaux -omiques et individus. Je suis intéressé par tester ce type d'approches, mais cela se fera dans un deuxième temps, car nous n'avons pas actuellement de données -omiques intermédiaires sur des populations suffisamment grandes. Pour ce faire, j'intégrerai l'acquisition de ce type de données dans mes futurs projets. Les études sur cet aspect s'attacheront aussi à certains aspects particulièrement importants chez des plantes pérennes, comme l'âge auquel collecter les données, la fréquence de collecte, le tissu, etc.



$$Y = \mu\mathbf{1} + Zu_G + Zu_T + Zu_P + Zu_M + \boldsymbol{\varepsilon}$$

$$u_G \sim MVN(0, G\sigma_g) \quad u_T \sim MVN(0, T\sigma_T)$$

$$u_P \sim MVN(0, P\sigma_P) \quad u_M \sim MVN(0, M\sigma_m)$$

$$\boldsymbol{\varepsilon} \sim MVN(0, I\sigma_\varepsilon^2)$$

Fig. 1 Biological hierarchy in genomic selection. Schematic of how various levels of the biological hierarchy of traits could be incorporated into GS models. The availability of the genomic (red), transcriptomic (orange), proteomic (green), and metabolomic (blue) data in maize makes it possible to incorporate multiple levels of the biological hierarchy of an agronomic trait directly into genomic selection (GS) models. Each of the different levels of the biological hierarchy can be used to calculate the correspondingly colored relationship matrices *G*, *T*, *P*, and *M*. Model terms and abbreviations: Y, observed vector trait values for *n* individuals; μ, grand mean; 1, *n*-dimensional vector of 1's; *u*, *n*-dimensional random vector of polygenic effects; ***Z***, incidence matrix relating u to Y; ε, *n*-dimensional random vector for error terms; MVN, multivariate normal

Figure 46 Les différents niveaux -omics et leur intégration dans un modèle de prédiction *multi-kernel* (Rice et Lipka, 2021)

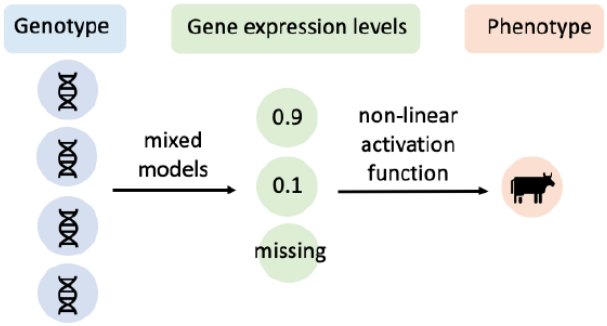


Figure 2 Framework of NN-LMM incorporating intermediate omics data such as gene expression levels. Genotypes affect the gene expression levels, then gene expression levels regulate the phenotypes. Linear mixed models can be applied to sample marker effects or genetic values on gene expression levels, and the nonlinear activation function in neural networks will be used to capture the complex nonlinear relationships between gene expression levels and phenotypes. For an individual, the gene expression levels of the first two genes are 0.9 and 0.1, respectively, and the gene expression of the last gene is missing to be sampled. Individuals can have different missing gene expression levels.

Figure 47 Schéma de principe du *neural networks linear mixed model* (Zhao et al., 2022)

3.4.4. Variants structuraux

L'utilisation de pangénomes est une autre voie possible pour une amélioration potentielle de la précision de la SG. Les progrès des techniques de séquençage ont permis de comparer les génomes individuels au sein des espèces et ont montré que les variations structurales (VS) représentaient une proportion importante du polymorphisme (Yuan et al., 2021). Les VS consistent en des délétions, insertions, variations du nombre de copies (CNV), inversions ou translocations, de taille >50 pb (Figure 42). En particulier, les VS incluent des variations dans la présence/absence de gènes, avec des gènes centraux qui sont présents chez tous les individus, et des gènes variables qui sont absents chez certains individus. Les VS ne peuvent pas être représentées par des génomes de référence uniques, et les pangénomes sont donc nécessaires pour exploiter l'ensemble de la diversité génétique de la population d'amélioration (Bayer et al., 2021 ; Scossa et al., 2021). Jusqu'à présent, très peu d'études ont envisagé d'utiliser les variations structurelles pour les prédictions génomiques. Chez le blé, Würschum et al. (2017a) ont obtenu une légère augmentation de la précision de la SG lorsque des marqueurs ciblant spécifiquement un CNV contribuant au contrôle génétique du caractère cible étaient inclus dans le modèle. De même, dans le maïs et le bétail, l'utilisation d'informations sur les CNV dans le modèle SG a augmenté la précision de la prédiction dans certains cas (Hay et al., 2018 ; Lyra et al., 2019). L'utilisation des informations de VS pour les prédictions génomiques mérite une plus grande attention, et cela sera grandement facilité par la disponibilité de pangénomes.

Plusieurs génomes de référence sont déjà disponibles chez le palmier à huile, et la prochaine étape devrait être la construction de pangénomes. Par ailleurs, le projet Cirad-PalmElit FreePalm porté par David Lopez va quantifier l'importance des VS chez les fondateurs du programme d'amélioration, ce qui sera très informatif sur l'intérêt d'approfondir cet aspect chez le palmier à huile.

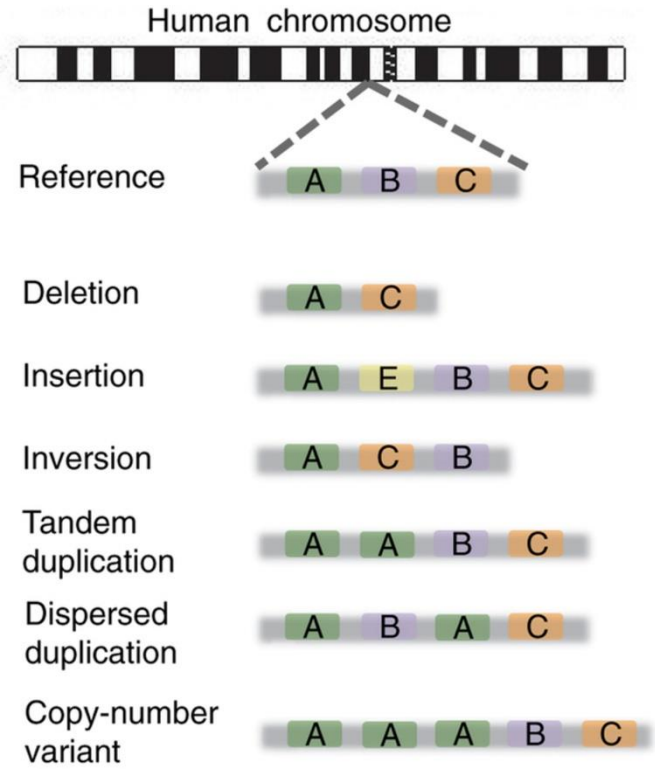


Figure 48 Illustration des différents types de variants structuraux
(https://www.gesundheitsindustrie-bw.de/en/article/news/genomic-structural-variations-can-cause-cancer)

3.5. Données multi-environnements

Les essais multi-environnements et les modèles de SG qui tiennent compte des effets environnementaux permettent de prédire la valeur génétique de nouveaux génotypes dans des environnements connus, de génotypes connus dans de nouveaux environnements et de nouveaux génotypes dans de nouveaux environnements (Bustos-Korts et al., 2016 ; Malosetti et al., 2016). La capacité à prédire les performances dans de nouveaux environnements présente un intérêt majeur dans le contexte du changement climatique, en particulier pour les cultures pérennes où la sélection souffre d'inertie en raison de la longueur des cycles de sélection. L'analyse des interactions entre le génotype et l'environnement (GxE) aide à sélectionner les génotypes qui sont stables à travers les environnements et ceux qui sont les meilleurs pour des environnements cibles spécifiques. La prise en compte des GxE dans les modèles de SG peut augmenter considérablement la précision de la prédiction lorsque des données provenant d'essais multi-environnements sont disponibles (Tong et Nikoloski, 2021 ; Xu et al., 2020). Diverses approches ont été développées pour intégrer les données environnementales dans les modèles de SG (Bustos-Korts et al., 2016 ; Crossa et al., 2017 ; Malosetti et al., 2016 ; Tong et Nikoloski, 2021 ; Xu et al., 2020). Les méthodes les plus intéressantes permettent de faire des prévisions dans de nouveaux environnements en utilisant des normes de réaction (Costa-Neto et Fritsche-Neto, 2021 ; Costa-Neto et al., 2021 ; Crossa et al., 2021) ou des modèles de croissance des cultures (*crop growth models*, CGM) (Crossa et al., 2021 ; Van Eeuwijk et al., 2019 ; Xu et al., 2020).

Les normes de réaction sont des fonctions linéaires ou non linéaires qui décrivent les phénotypes produits par un seul génotype à travers un gradient environnemental (Li et al., 2017). Elles peuvent être intégrées dans les analyses génétiques à l'aide de la régression aléatoire (Marchal et al., 2019 ; Mrode, 2014 ; Oliveira et al., 2019), ce qui conduit à des coefficients spécifiques au génotype qui caractérisent les normes de réaction pour chaque covariable environnementale. De manière équivalente, les covariables environnementales peuvent être utilisées pour construire une matrice d' « apparentement » environnemental qui quantifie les ressemblances entre les environnements considérés (Costa-Neto et al., 2021), à la manière des matrices d'apparentement calculées à partir des SNP.

Les CGM s'appuie sur les principes de la physiologie végétale, de la pédologie et de la climatologie pour modéliser le développement des plantes. Les CGM utilisent des équations impliquant des paramètres génétiques qui sont spécifiques aux génotypes considérés et sont supposés indépendants de l'environnement, et des variables environnementales (Boote et al., 2013). Plusieurs méthodes ont été développées pour intégrer les CGM dans le contexte de la SG (Crossa et al., 2021 ; Rincent et al., 2017). Le CGM peut être mis en œuvre pour prédire les stades de développement qui, avec les données météorologiques quotidiennes, seront utilisés pour calculer les covariables de stress climatique en fonction du stade de développement de la plante. Le CGM peut également être utilisé pour calculer les covariables de stress environnemental qui incluent la réponse de la culture aux conditions environnementales. Ces covariables environnementales peuvent ensuite être incorporées dans le modèle de SG en utilisant, par exemple, la régression aléatoire. On peut aussi estimer les paramètres génétiques du CGM pour les génotypes qui constituent la population de calibration et faire une prédiction génomique des paramètres génétiques des candidats à la sélection. L'utilisation du CGM et des covariables environnementales permet de prédire le phénotype des candidats à la sélection dans l'environnement cible. Cette approche a été qualifiée de *gene-based modeling*. Une autre méthode consiste à incorporer un CGM dans le modèle de prédiction génomique pour faire une estimation conjointe des effets des marqueurs et des paramètres génétiques du CGM. Cette méthode

est appelée CGM-WGP (*crop growth model-whole genome prediction*) et repose sur l'utilisation des méthodes statistiques *approximate Bayesian computation* ou de modèles hiérarchiques linéaires généralisés bayésiens.

Chez le palmier à huile, PalmElit et ses partenaires ont mis en place un réseau d'essais pour évaluer des croisements hybrides dans différents environnements (Bénin, Nigéria, Indonésie). J'ai d'ailleurs été impliqué dans ce travail lors de mon affectation au Bénin, notamment à l'étape de conception et de suivi du plan de croisements. Suffisamment de données phénotypiques seront bientôt disponibles pour permettre des analyses de SG, et la prise en compte des GxE utilisera les données fournies par les stations météorologiques des partenaires et/ou des bases de données publiques, telles WorldClim (http://www.worldclim.com/). Il sera aussi possible de s'appuyer sur les divers modèles de croissance déjà développés chez le palmier à huile, tels ECOPALM, APSIM-Oil Palm, CLM-Palm, CLIMEX-Oil Palm, PySawit, etc. (Teh Boon Sung et al., 2018).

3.6. Phénotypage haut-débit et sélection phénomique

Les plateformes de phénotypage à haut débit (*high-throughput phenotyping*, HTP) rendent le phénotypage plus rapide et donne des coûts de main-d'œuvre réduits par rapport aux méthodes conventionnelles (Persa et al., 2021). Le HTP permet (i) des analyses à l'échelle du champ avec des plateformes extérieures qui utilisent la télédétection et l'imagerie, principalement basées sur la spectroscopie dans le visible/le proche infrarouge et l'infrarouge lointain, et (ii) des analyses de la partie récoltable de la culture en utilisant la spectroscopie de réflectance dans le proche infrarouge (NIRS). Pour la SG, le HTP est un moyen efficace de caractériser de grandes populations de calibration (Wartha et Lorenz, 2021). Ceci est particulièrement utile pour les espèces pérennes qui nécessitent un phénotypage sur de longues périodes de temps. Le HTP a déjà été utilisé chez différentes plantes pérennes tropicales. Par exemple, les données multispectrales recueillies à partir d'un drone ont été utilisées pour estimer la hauteur et le diamètre chez l'eucalyptus (da Silva et al., 2021). Le NIRS a également été utilisée pour la quantification rapide des composants liés à la saveur du cacao et à la qualité du café Arabica (voir par exemple Álvarez et al. (2012); dos Santos Scholz et al. (2014)). Dans des populations d'eucalyptus utilisées pour la SG, le NIRS a été utilisé pour mesurer les caractères chimiques et physiques de la qualité du bois (Durán et al., 2017 ; de Moraes et al., 2018 ; Rambolarimanana et al., 2018).

Je souhaite utiliser le HTP pour évaluer l'efficacité de la sélection phénomique (Rincent et al., 2018), une méthode consistant à remplacer les données génomiques des modèles de SG par des données spectrales, ou à combiner les deux types de données. Dans la sélection phénomique, les différentes longueurs d'onde du spectre sont utilisées de la même façon que les marqueurs en SG, soit pour calculer des matrices d'apparentements entre individus qui serviront à implémenter des modèles de type GBLUP, soit pour implémenter directement des modèles de type RRBLUP (Brault et al., 2021 ; Persa et al., 2021 ; Rincent et al., 2018). La méthode repose sur l'idée que les différences entre individus en termes de profils spectraux reflètent des différences se situant à un niveau intermédiaire entre les génotypes et les phénotypes, c-à-d au niveau des différents endophénotypes (transcrits, métabolites, etc.) qui interviennent dans l'expression du caractère d'intérêt (Rincent et al., 2018). Cette utilisation du HTP est donc une autre façon d'exploiter des endophénotypes par rapport à ce que j'ai présenté dans la section 3.4.3. Elle est intéressante car elle permet d'accéder à un ensemble d'endophénotypes, sans *a priori* sur les mécanismes reliant génotype et phénotype. Par ailleurs, les

méthodes d'HTP, et en particulier le NIRS, permettent d'accéder à ces informations en haut-débit et à faible coût. La sélection phénomique a donné des résultats prometteurs, notamment chez des plantes pérennes. Par exemple, chez le peuplier, le gain génétique attendu avec la sélection phénomique était supérieur ou égal à celui de la SG, en fonction du caractère (Rincent et al., 2018). Chez l'eucalyptus, des modèle de prédiction utilisant des données spectrales ont donné des prédictions plus élevées que les modèles utilisant des données génomiques (Ballesta et al., 2022).

Chez le palmier à huile, nous possédons déjà des données de NIRS, mais elles ont été acquises sur les populations parentales et sur des nombres d'individus trop faibles pour une validation de la méthode. Différents chercheurs au Cirad sont intéressés par l'acquisition de données NIRS, pour des études qui ne se limitent pas à la sélection phénomique. Nous allons donc construire un projet qui permettra d'obtenir de telles données sur des individus hybrides évalués en essais et génotypés. Chez le teck, le workpackage dont je suis responsable dans le projet Bioteak inclus aussi une étude de sélection phénomique. Comme sur les prédictions multi-omiques, j'aurai à étudier certains aspects particulièrement importants chez les plantes pérennes, comme l'âge des individus au moment de l'acquisition des données NIRS, le tissu visé, etc.

Références

Abraham P. Stimulation of latex flow in Hevea brasiliensis of 4-amino-3, 5, 6-trichloropicolinic acid and 2-chloroethane-phosphonic acid. *J Rubber Res*. **1968**. Vol. 20, p. 291-305.

Absalome M. A., Massara C.-C., Alexandre A. A., Gervais K., Chantal G. G.-A., Ferdinand D., Rhedoor A. J., Coulibaly I., George T. G., Brigitte T., Marion M., Jean-Paul C. Biochemical properties, nutritional values, health benefits and sustainability of palm oil. *Biochimie*. **2020**. Vol. 178, p. 81-95. https://doi.org/10.1016/j.biochi.2020.09.019

Ahrends A., Hollingsworth P. M., Ziegler A. D., Fox J. M., Chen H., Su Y., Xu J. Current trends of rubber plantation expansion may threaten biodiversity and livelihoods. *Glob. Environ. Change*. **2015**. Vol. 34, p. 48-58. https://doi.org/https://doi.org/10.1016/j.gloenvcha.2015.06.002

Akdemir D., Isidro-Sánchez J. Design of training populations for selective phenotyping in genomic prediction. *Sci. Rep.* **2019**. Vol. 9, n°1, p. 1446. https://doi.org/10.1038/s41598-018-38081-6

Álvarez C., Pérez E., Cros E., Lares M., Assemat S., Boulanger R., Davrieux F. 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. Infrared Spectrosc.* **2012**. Vol. 20, n°2, p. 307-315.

Amiruddin N., Chan P.-L., Azizi N., Morris P. E., Chan K.-L., Ong P. W., Rosli R., Masura S. S., Murphy D. J., Sambanthamurthi R., Haslam R. P., Chye M.-L., Harwood J. L., Low E.-T. L. Characterization of Oil Palm Acyl-CoA-Binding Proteins and Correlation of Their Gene Expression with Oil Synthesis. *Plant Cell Physiol.* **2020**. Vol. 61, n°4, p. 735-747. https://doi.org/10.1093/pcp/pcz237

Babu K., Mathur R., Venu M., Shil S., Ravichandran G., Anita P., Bhagya H. Genome-wide association study (GWAS) of major QTLs for bunch and oil yield related traits in Elaeis guineensis L. *Plant Sci.* **2021**. Vol. 305, p. 110810.

Ballesta P., Ahmar S., Lobos G. A., Mieres-Castro D., Jiménez-Aspee F., Mora-Poblete F. Heritable Variation of Foliar Spectral Reflectance Enhances Genomic Prediction of Hydrogen Cyanide in a Genetically Structured Population of Eucalyptus. *Front. Plant Sci.* **2022**. Vol. 13, p. 871943. https://doi.org/10.3389/fpls.2022.871943

Ballesta P., Maldonado C., Pérez-Rodríguez P., Mora F. SNP and Haplotype-Based Genomic Selection of Quantitative Traits in Eucalyptus globulus. *Plants*. **2019**. Vol. 8, n°9, p. 331. https://doi.org/10.3390/plants8090331

Baumung R., Sölkner J., Essl A. Correlation between purebred and crossbred performance under a two-locus model with additive by additive interaction. *J. Anim. Breed. Genet.* **1997**. Vol. 114, n°1-6, p. 89-98. https://doi.org/10.1111/j.1439-0388.1997.tb00496.x

Bayer P. E., Petereit J., Danilevicz M. F., Anderson R., Batley J., Edwards D. The application of pangenomics and machine learning in genomic selection in plants. *Plant Genome*. **2021**. Vol. 14, n°3,. https://doi.org/10.1002/tpg2.20112 (consulté le 18 février 2022)

Beavis W. D. The power and deceit of QTL experiments: lessons from comparative QTL studies. *Proc. Forty-Ninth Annu. Corn Sorghum Ind. Res. Conf. Am. Seed Trade Assoc. Wash. DC*. **1994**. p. 250-266.

Beavis W. D. QTL analyses: Power, precision, and accuracy. In : *Mol. Dissection Complex Traits*. Boca Raton : Paterson, A.H., 1998. p. 145-162.

Beirnaert A., Vanderweyen R. Contribution à l'étude génétique et biométrique des variétiés d'*Elaeis guineensis* Jacq. *Publ Inst Nat Etude Agron Congo Belge Ser Sci*. **1941**. Vol. 27, p. 1-101.

Bejarano D., MartÃ\-nez R., Manrique C., Parra L. M., Rocha J. F., GÃ³mez Y., Abuabara Y., Gallego J. Linkage disequilibrium levels and allele frequency distribution in Blanco Orejinegro and Romosinuano Creole cattle using medium density SNP chip data. *Genet. Mol. Biol.* **2018**. Vol. 41, p. 426-433.

Bernardo R. Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. *Crop Sci.* **1994**. Vol. 34, n°1,. https://doi.org/10.2135/cropsci1994.0011183X003400010003x (consulté le 17 mars 2022)

Bernardo R. Reinventing quantitative genetics for plant breeding: something old, something new, something borrowed, something BLUE. *Heredity*. **2020**. Vol. 125, n°6, p. 375-385. https://doi.org/10.1038/s41437-020-0312-1

Bernardo R. Prospective Targeted Recombination and Genetic Gains for Quantitative Traits in Maize. *Plant Genome*. **2017**. Vol. 10, n°2,. https://doi.org/10.3835/plantgenome2016.11.0118 (consulté le 8 avril 2022)

Bernardo R. Genomewide Selection when Major Genes Are Known. *Crop Sci*. **2014**. Vol. 54, n°1, p. 68-75.

Bernhardsson C., Zan Y., Chen Z., Ingvarsson P. K., Wu H. X. Development of a highly efficient 50K single nucleotide polymorphism genotyping array for the large and complex genome of Norway spruce (Picea abies L. Karst) by whole genome resequencing and its transferability to other spruce species. *Mol Ecol Resour*. **2021**. Vol. 21, n°3, p. 880-896. https://doi.org/10.1111/1755-0998.13292

Bhat J. A., Ali S., Salgotra R. K., Mir Z. A., Dutta S., Jadon V., Tyagi A., Mushtaq M., Jain N., Singh P. K., Singh G. P., Prabhu K. V. Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. *Front. Genet.* **2016**. Vol. 7, p. 221. https://doi.org/10.3389/fgene.2016.00221

Bhat J. A., Yu D., Bohra A., Ganie S. A., Varshney R. K. Features and applications of haplotypes in crop breeding. *Commun. Biol.* **2021**. Vol. 4, n°1, p. 1266-1266. https://doi.org/10.1038/s42003-021-02782-y

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. QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis* Jacq.). *Theor. Appl. Genet.* **2010**. Vol. 120, n°8, p. 1673-1687. https://doi.org/10.1007/s00122-010-1284-y

Bini K., Saha T., Radhakrishnan S., Ravindran M., Uthup T. K. Development of Novel Markers for Yield in Hevea brasiliensis Muell. Arg. Based on Candidate Genes from Biosynthetic Pathways Associated with Latex Production. *Biochem. Genet.* **2022**.https://doi.org/10.1007/s10528-022-10211-w (consulté le 18 avril 2022)

Boote K. J., JONES J. W., WHITE J. W., ASSENG S., LIZASO J. I. Putting mechanisms into crop production models. *Plant Cell Environ.* **2013**. Vol. 36, n°9, p. 1658-1672. https://doi.org/10.1111/pce.12119

Brandariz S. P., Bernardo R. Predicted Genetic Gains from Targeted Recombination in Elite Biparental Maize Populations. *Plant Genome*. **2019**. Vol. 12, n°1, p. 180062. https://doi.org/10.3835/plantgenome2018.08.0062

Brault C., Lazerges J., Doligez A., Thomas M., Ecarnot M., Roumet P., Bertrand Y., Berger G., Pons T., François P., Le Cunff L., This P., Segura V. Interest of phenomic prediction as an alternative to genomic prediction in grapevine. **2021**.https://doi.org/10.1101/2021.12.16.472608 (consulté le 14 avril 2022)

Browning B. L., Tian X., Zhou Y., Browning S. R. Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **2021**. Vol. 108, n°10, p. 1880-1890. https://doi.org/10.1016/j.ajhg.2021.08.005

Browning B. L., Zhou Y., Browning S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **2018**. Vol. 103, n°3, p. 338-348. https://doi.org/10.1016/j.ajhg.2018.07.015

Bustos-Korts D., Malosetti M., Chapman S., Van Eeuwijk F. Modelling of Genotype by Environment Interaction and Prediction of Complex Traits across Multiple Environments as a Synthesis of Crop Growth Modelling, Genetics and Statistics. In : Yin X, Struik PC (éd.). *Crop Syst. Biol. Narrowing Gaps Crop Model. Genet.* Cham : Springer International Publishing, 2016. p. 55-82.https://doi.org/10.1007/978-3-319-20562-5_3ISBN : 978-3-319-20562-5.

Butler D. G., Cullis B. R., Gilmour A. R., Gogel B. J. *Mixed models for S language environments: ASReml-R reference manual (Version 3)*. [s.l.] : Queensland Department of Primary Industries and Fisheries, 2009. 398 p.

Caballero A. Developments in the prediction of effective population size. *Heredity*. **1994**. Vol. 73, n°6, p. 657-679.

Calleja-Rodriguez A., Pan J., Funda T., Chen Z., Baison J., Isik F., Abrahamsson S., Wu H. X. Evaluation of the efficiency of genomic versus pedigree predictions for growth and wood quality traits in Scots pine. *BMC Genomics*. **2020**. Vol. 21, n°1, p. 796. https://doi.org/10.1186/s12864-020-07188-4

Campbell M. T., Hu H., Yeats T. H., Brzozowski L. J., Caffe-Treml M., Gutiérrez L., Smith K. P., Sorrells M. E., Gore M. A., Jannink J.-L. Improving Genomic Prediction for Seed Quality Traits in Oat (Avena sativa L.) Using Trait-Specific Relationship Matrices. *Front. Genet.* **2021**. Vol. 12, p. 643733. https://doi.org/10.3389/fgene.2021.643733

De los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D., Calus M. P. L. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics*. **2013**. Vol. 193, n°2, p. 327-345. https://doi.org/10.1534/genetics.112.143313

De los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel K., Cotes J. M. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics*. **2009**. Vol. 182, n°1, p. 375-385. https://doi.org/10.1534/genetics.109.101501

Cantet R., Fernando R. Prediction of breeding values with additive animal models for crosses from 2 populations. *Genet. Sel. Evol.* **1995**. Vol. 27, n°4, p. 323. https://doi.org/10.1186/1297-9686-27-4-323

Cericola F., Lenk I., Fè D., Byrne S., Jensen C. S., Pedersen M. G., Asp T., Jensen J., Janss L. 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.* **2018**. Vol. 9, p. 369. https://doi.org/10.3389/fpls.2018.00369

Chagné D., Crowhurst R. N., Troggio M., Davey M. W., Gilmore B., Lawley C., Vanderzande S., Hellens R. P., Kumar S., Cestaro A., Velasco R., Main D., Rees J. D., Iezzoni A., Mockler T., Wilhelm L., Van de Weg E., Gardiner S. E., Bassil N., Peace C. Genome-Wide SNP Detection, Validation, and Development

of an 8K SNP Array for Apple. *PLoS One*. **2012**. Vol. 7, n°2, p. e31745. https://doi.org/10.1371/journal.pone.0031745

Chan K.-L., Tatarinova T. V., Rosli R., Amiruddin N., Azizi N., Halim M. A. A., Sanusi N. S. N. M., Jayanthi N., Ponomarenko P., Triska M., Solovyev V., Firdaus-Raih M., Sambanthamurthi R., Murphy D., Low E.-T. L. Evidence-based gene models for structural and functional annotations of the oil palm genome. *Biol. Direct*. **2017**. Vol. 12, n°1, p. 21. https://doi.org/10.1186/s13062-017-0191-4

Christensen O. F., Börner V., Varona L., Legarra A. Genetic evaluation including intermediate omics features. *Genetics*. **2021**. Vol. 219, n°2, p. iyab130. https://doi.org/10.1093/genetics/iyab130

Christensen O., Madsen P., Nielsen B., Su G. Genomic evaluation of both purebred and crossbred performances. *Genet. Sel. Evol.* **2014**. Vol. 46, n°1, p. 1-9. https://doi.org/10.1186/1297-9686-46-23

Clark S., Hickey J., Daetwyler H., Van der Werf J. 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.* **2012**. Vol. 44, n°1, p. 4.

Clément-Demange A., Priyadarshan P. M., Thuy Hoa T. T., Venkatachalam P. Hevea Rubber Breeding and Genetics. In : *Plant Breed. Rev.* [s.l.] : John Wiley & Sons, Inc., 2007. p. 177-283.https://doi.org/10.1002/9780470168035.ch4ISBN : 978-0-470-16803-5.

Cochard B. *Etude de la diversité génétique et du déséquilibre de liaison au sein de populations améliorées de palmier à huile (*Elaeis guineensis *Jacq.)*. Thèse de Doctorat. Montpellier : Montpellier SupAgro, 2008. 97-[175] p.

Cochard B., Adon B., Rekima S., Billotte N., De Chenon R., Koutou A., Nouy B., Omoré A., Purba A., Glazsmann J.-C., Noyer J.-L. Geographic and genetic structure of African oil palm diversity suggests new approaches to breeding. *Tree Genet. Genomes*. **2009**. Vol. 5, n°3, p. 493-504. https://doi.org/10.1007/s11295-009-0203-3

Combs E., Bernardo R. Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. *Plant Genome*. **2013**. Vol. 6, n°1,. https://doi.org/10.3835/plantgenome2012.11.0030 (consulté le 3 avril 2022)

Comstock R. E., Robinson H. F., Harvey P. H. A breeding procedure designed to make maximum use of both general and specific combining ability. *Agron J*. **1949**. Vol. 41, n°8, p. 360-367.

Conson A. R. O., Taniguti C. H., Amadeu R. R., Andreotti I. A. A., De Souza L. M., Dos Santos L. H. B., Rosa J. R. B. F., Mantello C. C., Da Silva C. C., José Scaloppi Junior E., Ribeiro R. V., Le Guen V., Garcia A. A. F., Gonçalves P. de S., De Souza A. P. High-Resolution Genetic Map and QTL Analysis of Growth-Related Traits of Hevea brasiliensis Cultivated Under Suboptimal Temperature and Humidity Conditions. *Front. Plant Sci.* **2018**. Vol. 9, p. 1255. https://doi.org/10.3389/fpls.2018.01255

Corley R. How much palm oil do we need ?. *Environ. Sci. Policy*. **2009**. Vol. 12, p. 134-139.

Corley R. H. V., Lee C. H. The physiological basis for genetic improvement of oil palm in Malaysia. *Euphytica*. **1992**. Vol. 60, n°3, p. 179-184. https://doi.org/10.1007/BF00039396

Corley R., Law I. The future for oil palm clones. In : *Proc Int Plant. Conf Incorp Soc Kuala Lumpur*. [s.l.] : [s.n.], 1997. p. 279-289.

Corley R., Tinker P. *The oil palm*. 5th éd. Chichester, UK : Wiley-Blackwell, 2016. 680 p.ISBN : 978-1-118-95329-7.

Costa-Neto G., Fritsche-Neto R. Enviromics: bridging different sources of data, building one framework. *Crop Breed. Appl. Biotechnol.* **2021**. Vol. 21,.

Costa-Neto G., Galli G., Carvalho H. F., Crossa J., Fritsche-Neto R. EnvRtype: a software to interplay enviromics and quantitative genomics in agriculture. *G3*. **2021**. Vol. 11, n°4, p. jkab040.

Coster A., Bastiaansen J. *HaploSim: R package version 1.8.4* [En ligne]. **2010**. Disponible sur : < http://CRAN.R-project.org/package=HaploSim >

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

Cros D., Bocs S., Riou V., Ortega-Abboud E., Tisné S., Argout X., Pomiès V., Nodichao L., Lubis Z., Cochard B., Durand-Gasselin T. Genomic preselection with genotyping-by-sequencing increases performance of commercial oil palm hybrid crosses. *BMC Genomics*. **2017**. Vol. 18, n°1, p. 839. https://doi.org/10.1186/s12864-017-4179-3

Cros D., Denis M., Bouvet J.-M., Sanchez L. Long-term genomic selection for heterosis without dominance in multiplicative traits: case study of bunch production in oil palm. *BMC Genomics*. **2015a**. Vol. 16, n°1, p. 651.

Cros D., Denis M., Sánchez L., Cochard B., Flori A., Durand-Gasselin T., Nouy B., Omoré A., Pomiès V., Riou V., Suryana E., Bouvet J.-M. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). *Theor. Appl. Genet.* **2015b**. Vol. 128, n°3, p. 397-410. https://doi.org/10.1007/s00122-014-2439-z

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. Within-family genomic selection in rubber tree (Hevea brasiliensis) increases genetic gain for rubber production. *Ind. Crops Prod.* **2019**. Vol. 138, p. 111464. https://doi.org/10.1016/j.indcrop.2019.111464

Cros D., Sánchez L., Cochard B., Samper P., Denis M., Bouvet J.-M., Fernández J. Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population. *Theor. Appl. Genet.* **2014**. Vol. 127, n°4, p. 981-994. https://doi.org/10.1007/s00122-014-2273-3

Cros D., Tchounke B., Nkague-Nkamba L. Training genomic selection models across several breeding cycles increases genetic gain in oil palm in silico study. *Mol. Breed.* **2018**. Vol. 38, n°7, p. 89. https://doi.org/10.1007/s11032-018-0850-x

Cros D., Tisné S., Nyouma A., Jacob F., Et al. Genotyping-by-sequencing and single nucleotide polymorphism array give similar prediction accuracies for genomic selection in oil palm (Elaeis guineensis Jacq.) hybrid breeding. **in prep**.

Crossa J., Fritsche-Neto R., Montesinos-Lopez O. A., Costa-Neto G., Dreisigacker S., Montesinos-Lopez A., Bentley A. R. The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. *Front. Plant Sci.* **2021**. Vol. 12,.

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. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* **2017**. Vol. 22, n°11, p. 961-975. https://doi.org/10.1016/j.tplants.2017.08.011

Daetwyler H. D., Calus M. P. L., Pong-Wong R., De los Campos G., Hickey J. M. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics*. **2013**. Vol. 193, n°2, p. 347-365. https://doi.org/10.1534/genetics.112.147983

Dai X., Yang X., Wang C., Fan Y., Xin S., Hua Y., Wang K., Huang H. CRISPR/Cas9-mediated genome editing in Hevea brasiliensis. *Ind. Crops Prod.* **2021**. Vol. 164, p. 113418. https://doi.org/10.1016/j.indcrop.2021.113418

Daval A., Pomiès V., Le Squin S., Denis M., Riou V., Breton F., Bink M., Cochard B., Jacob F., Billotte N., Others. In silico QTL mapping in an oil palm breeding program reveals a quantitative and complex genetic resistance to Ganoderma boninense. *Mol. Breed.* **2021**. Vol. 41, n°9, p. 1-18.

De Coster W., Weissensteiner M. H., Sedlazeck F. J. Towards population-scale long-read sequencing. *Nat. Rev. Genet.* **2021**. Vol. 22, n°9, p. 572-587. https://doi.org/10.1038/s41576-021-00367-3

Demol J., Baudoin J. P., Louant B. P., Maréchal R., Mergeai G., Otoul E. *Amélioration des plantes: Application aux principales espèces cultivées en régions tropicales*. Gembloux, Belgique : Presses Agronomiques de Gembloux, 2002. 581 p.

Denis M., Bouvet J.-M. Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. *Tree Genet. Genomes*. **2013**. Vol. 9, n°1, p. 37-51. https://doi.org/10.1007/s11295-012-0528-1

Van Dijk E. L., Auger H., Jaszczyszyn Y., Thermes C. Ten years of next-generation sequencing technology. *Trends Genet.* **2014**. Vol. 30, n°9, p. 418-426. https://doi.org/10.1016/j.tig.2014.07.001

Van Dijk E. L., Jaszczyszyn Y., Naquin D., Thermes C. The Third Revolution in Sequencing Technology. *Trends Genet.* **2018**. Vol. 34, n°9, p. 666-681. https://doi.org/10.1016/j.tig.2018.05.008

Domonhédo H., Cros D., Nodichao L., Billotte N., Ahanhanzo C. Enjeux et amélioration de la réduction de l'acidité dans les fruits mûrs du palmier à huile, *Elaeis guineensis* Jacq. (synthèse bibliographique). *Biotechnol. Agron. Société Environ.* **2018a**. Vol. 22, n°1,.

Domonhédo H., Cuéllar T., Espeout S., Droc G., Summo M., Rivallan R., Cros D., Nouy B., Omoré A., Nodichao L., Arondel V., Ahanhanzo C., Billotte N. Genomic structure, QTL mapping, and molecular markers of lipase genes responsible for palm oil acidity in the oil palm (Elaeis guineensis Jacq.). *Tree Genet. Genomes*. **2018b**. Vol. 14, n°5, p. 69. https://doi.org/10.1007/s11295-018-1284-7

Durán R., Isik F., Zapata-Valenzuela J., Balocchi C., Valenzuela S. Genomic predictions of breeding values in a cloned Eucalyptus globulus population in Chile. *Tree Genet. Genomes*. **2017**. Vol. 13, n°4, p. 74.

Durand-Gasselin T., Blangy L., Picasso C., De Franqueville H., Breton F., Amblard P., Cochard B., Louise C., Nouy B. Sélection du palmier à huile pour une huile de palme durable et responsabilité sociale. *OCL*. **2010**. Vol. 17, n°6, p. 385-392.

Edwards D., Batley J., Snowdon R. J. Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.* **2013**. Vol. 126, n°1, p. 1-11. https://doi.org/10.1007/s00122-012-1964-x

Elshire R. J., Glaubitz J. C., Sun Q., Poland J. A., Kawamoto K., Buckler E. S., Mitchell S. E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. **2011**. Vol. 6, n°5, p. e19379.

Endelman J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*. **2011**. Vol. 4,. https://doi.org/10.3835/plantgenome2011.08.0024

Esfandyari H., Sørensen A. C., Bijma P. A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genet. Sel. Evol.* **2015**. Vol. 47, n°1,. https://doi.org/10.1186/s12711-015-0155-z

Fan Y., Xin S., Dai X., Yang X., Huang H., Hua Y. Efficient genome editing of rubber tree (hevea brasiliensis) protoplasts using CRISPR/Cas9 ribonucleoproteins. *Ind. Crops Prod.* **2020**. Vol. 146, p. 112146. https://doi.org/10.1016/j.indcrop.2020.112146

FAO. *How to feed the world in 2050 ?* [En ligne]. [s.l.] : Food and Agriculture Organization of the United Nations, 2009. Disponible sur : < http://www.fao.org/fileadmin/templates/wsfs/docs/expert_paper/How_to_Feed_the_World_in_2050.pdf >

FAOSTAT. Crops and livestock products. [s.l.] : [s.n.], 2022. Disponible sur : < https://www.fao.org/faostat/en/#data/QCL > (consulté le 12 avril 2022)

Ferdosi M. H., Henshall J., Tier B. Study of the optimum haplotype length to build genomic relationship matrices. *Genet. Sel. Evol.* **2016**. Vol. 48, n°1, p. 75. https://doi.org/10.1186/s12711-016-0253-6

Feuillet C., Leach J. E., Rogers J., Schnable P. S., Eversole K. Crop genome sequencing: lessons and rationales. *Trends Plant Sci.* **2011**. Vol. 16, n°2, p. 77-88. https://doi.org/10.1016/j.tplants.2010.10.005

Fischer T., Byerlee D., Edmeades G. *Crop yields and global food security: will yield increase continue to feed the world?* Canberra : ACIAR, 2014. 634 p.(ACIAR monograph series, 158)ISBN : 978-1-925133-05-9.

Flint-Garcia S. A., Thornsberry J. M., Buckler E. S. Structure of Linkage Disequilibrium in Plants. *Annu. Rev. Plant Biol.* **2003**. Vol. 54, n°1, p. 357-374. https://doi.org/10.1146/annurev.arplant.54.031902.134907

Fugeray-Scarbel A., Bastien C., Dupont-Nivet M., Lemarié S., R2D2 Consortium. Why and how to switch to genomic selection: lessons from plant and animal breeding experience. *Front. Genet.* **2021**. Vol. 12, p. 1185.

Gallais A. *Théorie de la sélection en amélioration des plantes* [En ligne]. [s.l.] : Masson, 1990. 588 p.(Collection Sciences agronomiques). Disponible sur : < http://books.google.fr/books?id=mjUhAQAAMAAJ >ISBN : 978-2-225-81424-2.

Gallais A. *Hétérosis et variétés hybrides en amélioration des plantes*. Versailles, France : Quae éditions, 2009. 376 p.(Synthèses). ISBN : 978-2-7592-0374-1.

Garcia D. *Amélioration génétique de l'hévéa pour la résistance au SALB et la production de latex. Etudes génomiques et génétiques des résistances de l'hévéa à Pseudocercospora ulei*. Mémoire présenté pour l'obtention de l'Habilitation à Diriger des Recherches. [s.l.] : Université de Montpellier, 2017.

Garrick D., Dekkers J., Fernando R. The evolution of methodologies for genomic prediction. *Livest. Sci.* **2014**. Vol. 166, p. 10-18. https://doi.org/10.1016/j.livsci.2014.05.031

Gascon J. P., De Berchoux C. Caractéristique de la production d'Elaeis guineensis (Jacq.) de diverses origines et de leurs croisements - Application à la sélection du palmier à huile. *Oléagineux*. **1964**. Vol. 19, n°2, p. 75-84.

Gascon J. P., Noiret J. M., Bénard G. Contribution à l'étude de l'hérédité de la production de régimes d'*Elaeis guineensis* Jacq. - Application à la sélection du palmier à huile. *Oléagineux*. **1966**. Vol. 21, n°11, p. 657-661.

Gaynor R. C., Gorjanc G., Hickey J. M. AlphaSimR: an R package for breeding program simulations. *G3 GenesGenomesGenetics*. **2021**. Vol. 11, n°jkaa017,. https://doi.org/10.1093/g3journal/jkaa017 (consulté le 21 mars 2021)

Gianola D., Van Kaam J. B. C. H. M. Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. *Genetics*. **2008**. Vol. 178, n°4, p. 2289-2303. https://doi.org/10.1534/genetics.107.084285

Glaubitz J. C., Casstevens T. M., Lu F., Harriman J., Elshire R. J., Sun Q., Buckler E. S. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLoS One*. **2014**. Vol. 9, n°2, p. e90346. https://doi.org/10.1371/journal.pone.0090346

Goddard M. E., Hayes B. J. Genomic selection. *J Anim Breed Genet*. **2007**. Vol. 124,. https://doi.org/10.1111/j.1439-0388.2007.00702.x

Goddard M. E., Kemper K. E., MacLeod I. M., Chamberlain A. J., Hayes B. J. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proc. Biol. Sci.* **2016**. Vol. 283, n°1835, p. 20160569. https://doi.org/10.1098/rspb.2016.0569

González-Diéguez D., Legarra A., Charcosset A., Moreau L., Lehermeier C., Teyssèdre S., Vitezica Z. G. Genomic prediction of hybrid crops allows disentangling dominance and epistasis. *Genetics*. **2021**. Vol. 218, n°1, p. iyab026. https://doi.org/10.1093/genetics/iyab026

Goodyear C. *Gum-elastic and its Varieties: With a Detailed Account of its Applications and Uses, and of the Discovery of Vulcanization*. [s.l.] : Published for the author, 1853.

Gorjanc G., Hickey J. M. AlphaMate: a program for optimizing selection, maintenance of diversity and mate allocation in breeding programs. *Bioinformatics*. **2018**. Vol. 34, n°19, p. 3408-3411.

Grafton R. Q., Williams J., Jiang Q. Food and water gaps to 2050: preliminary results from the global food and water system (GFWS) platform. *Food Secur.* **2015**. Vol. 7, n°2, p. 209-220. https://doi.org/10.1007/s12571-015-0439-8

Grattapaglia D. Breeding forest trees by genomic selection: current progress and the way forward. In : *Genomics Plant Genet. Resour.* [s.l.] : Tuberosa R, Graner A, Frison E, 2014. p. 651-682.

Grattapaglia D. Status and Perspectives of Genomic Selection in Forest Tree Breeding. In : Varshney RK, Roorkiwal M, Sorrells ME (éd.). *Genomic Sel. Crop Improv. New Mol. Breed. Strateg. Crop Improv.* Cham : Springer International Publishing, 2017. p. 199-249.https://doi.org/10.1007/978-3-319-63170-7_9ISBN : 978-3-319-63170-7.

Grattapaglia D., Resende M. D. V. Genomic selection in forest tree breeding. *Tree Genet Genomes*. **2011**. Vol. 7, p. 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. Quantitative Genetics and Genomics Converge to Accelerate Forest Tree Breeding. *Front. Plant Sci.* **2018**. Vol. 9, p. 1693. https://doi.org/10.3389/fpls.2018.01693

Gupta P. K., Kumar J., Mir R. R., Kumar A. Marker-Assisted Selection as a Component of Conventional Plant Breeding. In : *Plant Breed. Rev.* [s.l.] : John Wiley & Sons, Ltd, 2010. p. 145-217.https://doi.org/10.1002/9780470535486.ch4 (consulté le 19 février 2021)ISBN : 978-0-470-53548-6.

Gupta P. K., Rustgi S., Kulwal P. L. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. *Plant Mol. Biol.* **2005**. Vol. 57, n°4, p. 461-485. https://doi.org/10.1007/s11103-005-0257-z

Gutiérrez J. P., Goyache F. A note on ENDOG: a computer program for analysing pedigree information. *J. Anim. Breed. Genet.* **2005**. Vol. 122, n°3, p. 172-176. https://doi.org/10.1111/j.1439-0388.2005.00512.x

Habier D., Fernando R., Kizilkaya K., Garrick D. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*. **2011**. Vol. 12, n°1, p. 186.

Habier D., Fernando R. L., Dekkers J. C. M. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*. **2007**. Vol. 177, n°4, p. 2389-2397. https://doi.org/10.1534/genetics.107.081190

Hamilton M. B. *Population genetics*. Second edition. Hoboken, NJ : Wiley-Blackwell, 2021. ISBN : 978-1-118-43694-3.

Hay E. H. A., Utsunomiya Y. T., Xu L., Zhou Y., Neves H. H. R., Carvalheiro R., Bickhart D. M., Ma L., Garcia J. F., Liu G. E. Genomic predictions combining SNP markers and copy number variations in Nellore cattle. *BMC Genomics*. **2018**. Vol. 19, n°1, p. 441. https://doi.org/10.1186/s12864-018-4787-6

Hayes B. J., Bowman P. J., Chamberlain A. J., Goddard M. E. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* **2009**. Vol. 92, n°2, p. 433-443. https://doi.org/10.3168/jds.2008-1646

Hazel L. N., Lush J. L. THE EFFICIENCY OF THREE METHODS OF SELECTION*. *J. Hered.* **1942**. Vol. 33, n°11, p. 393-399. https://doi.org/10.1093/oxfordjournals.jhered.a105102

Heffner E. L., Sorrells M. E., Jannink J.-L. Genomic selection for crop improvement. *Crop Sci*. **2009**. Vol. 49, n°1, p. 1-12.

Henderson C. R. Estimation of genetic parameters. *Ann Math Stat.* **1950**. Vol. 21, p. 309-310.

Henderson C. R. *Applications of linear models in animal breeding*. [s.l.] : University of Guelph, 1984.

Henderson C. R. Statistical methods in animal improvement: Historical Overview. In : *Adv. Stat. Methods Genet. Improv. Lifestock*. Berlin : Gianola D, Hammond K, 1986. p. 2-14.

Heslot N., Jannink J.-L., Sorrells M. E. Perspectives for Genomic Selection Applications and Research in Plants. *Crop Sci.* **2015**. Vol. 55, n°1, p. 1-12. https://doi.org/10.2135/cropsci2014.03.0249

Hess M., Druet T., Hess A., Garrick D. Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genet. Sel. Evol.* **2017**. Vol. 49, n°1, p. 54. https://doi.org/10.1186/s12711-017-0329-y

Hickey J., Kinghorn B., Tier B., Clark S. A., Van der Werf J., Gorjanc G. Genomic evaluations using similarity between haplotypes. *J. Anim. Breed. Genet.* **2013**. Vol. 130, n°4, p. 259-269.

Hoban S., Bertorelle G., Gaggiotti O. E. Computer simulations: tools for population and evolutionary genetics. *Nat. Rev. Genet.* **2012**. Vol. 13, n°2, p. 110-122. https://doi.org/10.1038/nrg3130

Hu T., Chitnis N., Monos D., Dinh A. Next-generation sequencing technologies: An overview. *Hum. Immunol.* **2021**. Vol. 82, n°11, p. 801-811. https://doi.org/10.1016/j.humimm.2021.02.012

Ibánez-Escriche N., Fernando R., Toosi A., Dekkers J. Genomic selection of purebreds for crossbred performance. *Genet. Sel. Evol.* **2009**. Vol. 41, n°1, p. 12.

Isidro J., Jannink J.-L., Akdemir D., Poland J., Heslot N., Sorrells MarkE. Training set optimization under population structure in genomic selection. *Theor. Appl. Genet.* **2015**. Vol. 128, n°1, p. 145-158. https://doi.org/10.1007/s00122-014-2418-4

Isidro y Sánchez J., Akdemir D. Training Set Optimization for Sparse Phenotyping in Genomic Selection: A Conceptual Overview. *Front. Plant Sci.* **2021**. Vol. 12, p. 715910. https://doi.org/10.3389/fpls.2021.715910

Isik F. Genomic selection in forest tree breeding: the concept and an outlook to the future. *New For.* **2014**. Vol. 45, n°3, p. 379-401. https://doi.org/10.1007/s11056-014-9422-z

Ithnin M., Xu Y., Marjuni M., Serdari N. M., Amiruddin M. D., Low E.-T. L., Tan Y.-C., Yap S.-J., Ooi L. C. L., Nookiah R., Singh R., Xu S. Multiple locus genome-wide association studies for important economic traits of oil palm. *Tree Genet. Genomes*. **2017**. Vol. 13, n°5, p. 103. https://doi.org/10.1007/s11295-017-1185-1

Jin J., Lee M., Bai B., Sun Y., Qu J., Rahmadsyah, Alfiko Y., Lim C. H., Suwanto A., Sugiharti M., Wong L., Ye J., Chua N.-H., Yue G. H. Draft genome sequence of an elite *Dura* palm and whole-genome patterns of DNA variation in oil palm. *DNA Res.* **2016**. Vol. 23, n°6, p. 527-533. https://doi.org/10.1093/dnares/dsw036

Kinghorn B. P., Hickey J. M., Van Der Werf J. H. J. Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals. In : *Proc. 9th World Congr. Genet. Appl. Livest. Prod.* Leipzig, Germany : [s.n.], 2010. p. 36.

Knoch D., Werner C. R., Meyer R. C., Riewe D., Abbadi A., Lücke S., Snowdon R. J., Altmann T. Multi-omics-based prediction of hybrid performance in canola. *Theor. Appl. Genet.* **2021**. Vol. 134, n°4, p. 1147-1165. https://doi.org/10.1007/s00122-020-03759-x

De Koning D.-J. Meuwissen *et al.* on Genomic Selection. *Genetics*. **2016**. Vol. 203, n°1, p. 5-7. https://doi.org/10.1534/genetics.116.189795

Kumar K. R., Cowley M. J., Davis R. L. Next-Generation Sequencing and Emerging Technologies. *Semin. Thromb. Hemost.* **2019**. Vol. 45, n°07, p. 661-673. https://doi.org/10.1055/s-0039-1688446

Kumar S., Molloy C., Muñoz P., Daetwyler H., Chagné D., Volz R. Genome-Enabled Estimates of Additive and Non-additive Genetic Variances and Prediction of Apple Phenotypes Across Environments. *G3 GenesGenomesGenetics*. **2015**.https://doi.org/10.1534/g3.115.021105

Kwong Q. B., Ong A. L., Teh C. K., Chew F. T., Tammi M., Mayes S., Kulaveerasingam H., Yeoh S. H., Harikrishna J. A., Appleton D. R. Genomic Selection in Commercial Perennial Crops: Applicability and Improvement in Oil Palm (Elaeis guineensis Jacq.). *Sci. Rep.* **2017a**. Vol. 7, n°1, p. 2872. https://doi.org/10.1038/s41598-017-02602-6

Kwong Q. B., Teh C. K., Ong A. L., Chew F. T., Mayes S., Kulaveerasingam H., Tammi M., Yeoh S. H., Appleton D. R., Harikrishna J. A. Evaluation of methods and marker Systems in Genomic Selection of oil palm (Elaeis guineensis Jacq.). *BMC Genet.* **2017b**. Vol. 18, n°1, p. 107.

Kwong Q. B., Teh C. K., Ong A. L., Heng H. Y., Lee H. L., Mohamed M., Low J. Z.-B., Apparow S., Chew F. T., Mayes S., Kulaveerasingam H., Tammi M., Appleton D. R. Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm. *Mol. Plant*. **2016**. Vol. 9, n°8, p. 1132-1141. https://doi.org/10.1016/j.molp.2016.04.010

Lande R., Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*. **1990**. Vol. 124, n°3, p. 743-756.

Le Guen V., Doaré F., Weber C., Seguin M. Genetic structure of Amazonian populations of *Hevea brasiliensis* is shaped by hydrographical network and isolation by distance. *Tree Genet. Genomes*. **2009**. Vol. 5, n°4, p. 673-683. https://doi.org/10.1007/s11295-009-0218-9

Le Mouël C., Forslund A. How can we feed the world in 2050? A review of the responses from global scenario studies. *Eur. Rev. Agric. Econ.* **2017**. Vol. 44, n°4, p. 541-591. https://doi.org/10.1093/erae/jbx006

Lebedev V. G., Lebedeva T. N., Chernodubov A. I., Shestibratov K. A. Genomic Selection for Forest Tree Improvement: Methods, Achievements and Perspectives. *Forests*. **2020**. Vol. 11, n°11, p. 1190. https://doi.org/10.3390/f11111190

Lecerf J.-M. L'huile de palme. *Médecine Mal. Métaboliques*. **2017**. Vol. 11, n°4, p. 347-352. https://doi.org/10.1016/S1957-2557(17)30079-2

Legarra A., Robert-Granie C., Manfredi E., Elsen J. M. Performance of genomic selection in mice. *Genetics*. **2008**. Vol. 180, p. 611-618.

Lenz P. R., Beaulieu J., Mansfield S. D., Clément S., Desponts M., Bousquet J. 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*. **2017**. Vol. 18, n°1, p. 335.

Li K., Tscharntke T., Saintes B., Buchori D., Grass I. Critical factors limiting pollination success in oil palm: A systematic review. *Agric. Ecosyst. Environ.* **2019**. Vol. 280, p. 152-160. https://doi.org/10.1016/j.agee.2019.05.001

Li Y., Suontama M., Burdon R. D., Dungey H. S. Genotype by environment interactions in forest tree breeding: review of methodology and perspectives on research and application. *Tree Genet. Genomes*. **2017**. Vol. 13, n°3, p. 1-18.

Lin Z., Hayes B. J., Daetwyler H. D. Genomic selection in crops, trees and forages: a review. *Crop Pasture Sci.* **2014**. Vol. 65, n°11, p. 1177-1191.

Liu J., Shi Cong, Shi C.-C., Li W., Zhang Q.-J., Zhang Y., Li K., Lu H.-F., Shi Chao, Zhu S.-T., Xiao Z.-Y., Nan H., Yue Y., Zhu X.-G., Wu Y., Hong X.-N., Fan G.-Y., Tong Y., Zhang D., Mao C.-L., Liu Y.-L., Hao S.-J., Liu W.-Q., Lv M.-Q., Zhang H.-B., Liu Y., Hu-Tang G.-R., Wang J.-P., Wang J.-H., Sun Y.-H., Ni S.-B., Chen W.-B., Zhang X.-C., Jiao Y.-N., Eichler E. E., Li G.-H., Liu X., Gao L.-Z. The Chromosome-Based Rubber Tree Genome Provides New Insights into Spurge Genome Evolution and Rubber Biosynthesis. *Mol. Plant*. **2020**. Vol. 13, n°2, p. 336-350. https://doi.org/10.1016/j.molp.2019.10.017

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. Factors affecting genomic selection revealed by empirical evidence in maize. *Crop J.* **2018**. Vol. 6, n°4, p. 341-352. https://doi.org/10.1016/j.cj.2018.03.005

Lo L. L., Fernando R. L., Grossman M. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. *J. Anim. Sci.* **1997**. Vol. 75, n°11, p. 2877-2884.

Long H. S., Powell G., Greenaway S., Mallon A.-M., Lindgren C. M., Simon M. M. Making sense of the linear genome, gene function and TADs. *bioRxiv*. **2020**. p. 2020.09.28.316786. https://doi.org/10.1101/2020.09.28.316786

Long N., Gianola D., Rosa G. J. M., Weigel K. A. Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor. Appl. Genet.* **2011**. Vol. 123, n°7, p. 1065. https://doi.org/10.1007/s00122-011-1648-y

Long X., Fang Y., Qin Y., Yang J., Xiao X. Latex-specific transcriptome analysis reveals mechanisms for latex metabolism and natural rubber biosynthesis in laticifers of Hevea brasiliensis. *Ind. Crops Prod.* **2021**. Vol. 171, p. 113835. https://doi.org/10.1016/j.indcrop.2021.113835

Lorenz A. J., Chao S., Asoro F. G., Heffner E. L., Hayashi T., Iwata H., Smith K. P., Sorrells M. E., Jannink J.-L. Genomic Selection in Plant Breeding: Knowledge and Prospects. In : Donald L. Sparks (éd.). *Adv. Agron.* [s.l.] : Academic Press, 2011. p. 77-123. Disponible sur : < http://www.sciencedirect.com/science/article/pii/B9780123855312000025 >ISBN : 0065-2113.

Lourenco D., Legarra A., Tsuruta S., Masuda Y., Aguilar I., Misztal I. Single-Step Genomic Evaluations from Theory to Practice: Using SNP Chips and Sequence Data in BLUPF90. *Genes*. **2020**. Vol. 11, n°7, p. 790. https://doi.org/10.3390/genes11070790

Lynch M. Estimation of relatedness by DNA fingerprinting. *Mol Biol Evol*. **1988**. Vol. 5, n°5, p. 584-599.

Lynch M., Walsh B. *Genetics and analysis of quantitative traits*. Sunderland, MA : Sinauer Associates, Inc., 1998. 980 p.

Lyra D. H., Galli G., Alves F. C., Granato Í. S. C., Vidotti M. S., Bandeira e Sousa M., Morosini J. S., Crossa J., Fritsche-Neto R. Modeling copy number variation in the genomic prediction of maize hybrids. *Theor. Appl. Genet.* **2019**. Vol. 132, n°1, p. 273-288. https://doi.org/10.1007/s00122-018-3215-2

Mackay I., Powell W. Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci.* **2007**. Vol. 12, n°2, p. 57-63. https://doi.org/10.1016/j.tplants.2006.12.001

MacLeod I. M., Bowman P. J., Vander Jagt C. J., Haile-Mariam M., Kemper K. E., Chamberlain A. J., Schrooten C., Hayes B. J., Goddard M. E. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. *BMC Genomics*. **2016**. Vol. 17, n°1, p. 144. https://doi.org/10.1186/s12864-016-2443-6

Maldonado C., Mora-Poblete F., Contreras-Soto R. I., Ahmar S., Chen J.-T., Do Amaral Júnior A. T., Scapim C. A. Genome-Wide Prediction of Complex Traits in Two Outcrossing Plant Species Through Deep Learning and Bayesian Regularized Neural Network. *Front. Plant Sci.* **2020**. Vol. 11, p. 593897. https://doi.org/10.3389/fpls.2020.593897

Malosetti M., Bustos-Korts D., Boer M. P., Van Eeuwijk F. A. Predicting responses in multiple environments: issues in relation to genotype$\times$ environment interactions. *Crop Sci.* **2016**. Vol. 56, n°5, p. 2210-2222.

Marchal A., Legarra A., Tisné S., Carasco-Lacombe C., Manez A., Suryana E., Omoré A., Durand-Gasselin T., Sánchez L., Bouvet J.-M., Cros D. Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests. *Mol. Breed.* **2016**. Vol. 36, n°2, p. 1-13. https://doi.org/10.1007/s11032-015-0423-1

Marchal A., Schlichting C. D., Gobin R., Balandier P., Millier F., Muñoz F., Pâques L. E., Sánchez L. Deciphering hybrid larch reaction norms using random regression. *G3 Genes Genomes Genet.* **2019**. Vol. 9, n°1, p. 21-32.

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. Prediction of Cacao (Theobroma cacao) Resistance to Moniliophthora spp. Diseases via Genome-Wide Association Analysis and Genomic Selection. *Front. Plant Sci.* **2018**. Vol. 9, p. 343. https://doi.org/10.3389/fpls.2018.00343

Merrick L. F., Herr A. W., Sandhu K. S., Lozada D. N., Carter A. H. Optimizing Plant Breeding Programs for Genomic Selection. *Agronomy*. **2022**. Vol. 12, n°3, p. 714. https://doi.org/10.3390/agronomy12030714

Meunier J., Gascon J. Le schéma général d'amélioration du palmier à huile à l'IRHO. *Oléagineux*. **1972**. Vol. 27, n°1, p. 1-12.

Meunier J., Gascon J. P., Noiret J. M. Hérédité des caractéristiques du régime d'*Elaeis guineensis* Jacq. en Côte d'Ivoire. *Oléagineux*. **1970**. Vol. 25, p. 377-382.

Meuwissen T. H. E., Hayes B. J., Goddard M. E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*. **2001**. Vol. 157, n°4, p. 1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Michel S., Kummer C., Gallee M., Hellinger J., Ametz C., Akgöl B., Epure D., Löschenberger F., Buerstmayr H. Improving the baking quality of bread wheat by genomic selection in early generations. *Theor. Appl. Genet.* **2018**. Vol. 131, n°2, p. 477-493. https://doi.org/10.1007/s00122-017-2998-x

Minamikawa M. F., Nonaka K., Kaminuma E., Kajiya-Kanegae H., Onogi A., Goto S., Yoshioka T., Imai A., Hamada H., Hayashi T., Matsumoto S., Katayose Y., Toyoda A., Fujiyama A., Nakamura Y., Shimizu T., Iwata H. Genome-wide association study and genomic prediction in citrus: Potential of genomics-assisted breeding for fruit quality traits. *Sci. Rep.* **2017**. Vol. 7, n°1, p. 4721. https://doi.org/10.1038/s41598-017-05100-x

Momen M., Mehrgardi A. A., Sheikhi A., Kranis A., Tusell L., Morota G., Rosa G. J. M., Gianola D. Predictive ability of genome-assisted statistical models under various forms of gene action. *Sci. Rep.* **2018**. Vol. 8, n°1, p. 12309. https://doi.org/10.1038/s41598-018-30089-2

Money D., Migicovsky Z., Gardner K., Myles S. LinkImputeR: user-guided genotype calling and imputation for non-model organisms. *BMC Genomics*. **2017**. Vol. 18, n°1, p. 523. https://doi.org/10.1186/s12864-017-3873-5

Montesinos López O. A., Montesinos López A., Crossa J. *Multivariate Statistical Machine Learning Methods for Genomic Prediction* [En ligne]. Cham : Springer International Publishing, 2022. https://doi.org/10.1007/978-3-030-89010-0 (consulté le 7 avril 2022)ISBN : 978-3-030-89009-4.

Montesinos-López O. A., Montesinos-López A., Pérez-Rodríguez P., Barrón-López J. A., Martini J. W. R., Fajardo-Flores S. B., Gaytan-Lugo L. S., Santana-Mancilla P. C., Crossa J. A review of deep learning applications for genomic selection. *BMC Genomics*. **2021**. Vol. 22, n°1, p. 19. https://doi.org/10.1186/s12864-020-07319-x

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., Resende M. F. R., Muñoz P. R. Genomic selection prediction models comparing sequence capture and SNP array genotyping methods. *Mol. Breed.* **2018**. Vol. 38, n°9, p. 115. https://doi.org/10.1007/s11032-018-0865-3

Morota G., Gianola D. Kernel-based whole-genome prediction of complex traits: a review. *Front. Genet.* **2014**. Vol. 5, p. 363.

Mota-Gómez I., Lupiáñez D. G. A (3D-Nuclear) Space Odyssey: Making Sense of Hi-C Maps. *Genes*. **2019**. Vol. 10, n°6, p. 415. https://doi.org/10.3390/genes10060415

Mrode R. A. *Linear models for the prediction of animal breeding values*. 3e éd. Boston, MA : CABI, 2014. 343 p.

Mrode R. A. *Linear models for the prediction of animal breeding values*. 2e éd. Oxfordshire, UK : CABI, 2005. 344 p.

Mrode R., Ojango J. M. K., Okeyo A. M., Mwacharo J. M. 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.* **2019**. Vol. 9,. https://doi.org/10.3389/fgene.2018.00694 (consulté le 6 juillet 2020)

Munyengwa N., Le Guen V., Bille H. N., Souza L. M., Clément-Demange A., Mournet P., Masson A., Soumahoro M., Kouassi D., Cros D. 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*. **2021**. Vol. 113, n°2, p. 655-668. https://doi.org/10.1016/j.ygeno.2021.01.012

Muranty H., Jorge V., Bastien C., Lepoittevin C., Bouffier L., Sanchez L. Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. *Tree Genet. Genomes*. **2014**. p. 1-20. https://doi.org/10.1007/s11295-014-0790-5

Muranty H., Troggio M., Sadok I. B., Rifaï M. A., Auwerkerken A., Banchi E., Velasco R., Stevanato P., Van de Weg W. E., Di Guardo M., Kumar S., Laurens F., Bink M. C. A. M. Accuracy and responses of genomic selection on key traits in apple breeding. *Hortic. Res.* **2015**. Vol. 2, p. 15060. https://doi.org/10.1038/hortres.2015.60

Murphy D. J. The Future of Oil Palm as a Major Global Crop: Opportunities and Challenges. *J. Oil Palm Res.* **2014**. Vol. 26, n°1, p. 1-24.

Myles S. Improving fruit and wine: what does genomics have to offer?. *Trends Genet.* **2013**. Vol. 29, n°4, p. 190-196. https://doi.org/10.1016/j.tig.2013.01.006

Nakano Y., Mitsuda N., Ide K., Mori T., Mira F. R., Rosmalawati S., Watanabe N., Suzuki K. Transcriptome analysis of Pará rubber tree (H. brasiliensis) seedlings under ethylene stimulation. *BMC Plant Biol.* **2021**. Vol. 21, n°1, p. 420. https://doi.org/10.1186/s12870-021-03196-y

Nielsen N. H., Jahoor A., Jensen J. D., Orabi J., Cericola F., Edriss V., Jensen J. Genomic Prediction of Seed Quality Traits Using Advanced Barley Breeding Lines. *PLOS ONE*. **2016**. Vol. 11, n°10, p. e0164494. https://doi.org/10.1371/journal.pone.0164494

Noel S., Mikulcak F., Etter H., Stewart N. *Economics of Land Degradation Initiative: Report for policy and decision makers - Reaping economic and environmental benefits from sustainable land management* [En ligne]. Bonn, Germany : ELD Initiative and Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH, 2015. Disponible sur : < https://repo.mel.cgiar.org/handle/20.500.11766/4881 >

Nouy B., Jacquemard J.-C., Suryana E., Potier F., Konan K. E., Durand-Gasselin T. The expected and observed characteristics of several oil palm (*Elaeis guineensis* Jacq.) clones. In : IOPRI (éd.). *International Oil Palm Conference*. public : s.n., 2006. p. 17 p.

Nsibi M., Gouble B., Bureau S., Flutre T., Sauvage C., Audergon J.-M., Regnard J.-L. Adoption and Optimization of Genomic Selection To Sustain Breeding for Apricot Fruit Quality. *G3 Bethesda Md*. **2020**. Vol. 10, n°12, p. 4513-4529. https://doi.org/10.1534/g3.120.401452

Nyine M., Uwimana B., Blavet N., Hřibová E., Vanrespaille H., Batte M., Akech V., Brown A., Lorenzen J., Swennen R., Doležel J. Genomic Prediction in a Multiploid Crop: Genotype by Environment Interaction and Allele Dosage Effects on Predictive Ability in Banana. *Plant Genome*. **2018**. Vol. 11, n°2, p. 170090. https://doi.org/10.3835/plantgenome2017.10.0090

Nyouma A., Bell J. M., Jacob F., Cros D. From mass selection to genomic selection: one century of breeding for quantitative yield components of oil palm (Elaeis guineensis Jacq.). *Tree Genet. Genomes*. **2019**. Vol. 15, n°5, p. 69. https://doi.org/10.1007/s11295-019-1373-2

Nyouma A., Bell J. M., Jacob F., Riou V., Manez A., Pomiès V., Domonhedo H., Arifiyanto D., Cochard B., Durand-Gasselin T., Cros D. 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*. **2022**.https://doi.org/10.1007/s00438-022-01867-5

Nyouma A., Bell J. M., Jacob F., Riou V., Manez A., Pomiès V., Nodichao L., Syahputra I., Affandi D., Cochard B., Durand-Gasselin T., Cros D. Genomic predictions improve clonal selection in oil palm (Elaeis guineensis Jacq.) hybrids. *Plant Sci.* **2020**. Vol. 299, p. 110547. https://doi.org/10.1016/j.plantsci.2020.110547

Oliveira H. R., Brito L. F., Lourenco D. A. L., Silva F. F., Jamrozik J., Schaeffer L. R., Schenkel F. S. Invited review: Advances and applications of random regression models: From quantitative genetics to genomics. *J. Dairy Sci.* **2019**. Vol. 102, n°9, p. 7664-7683. https://doi.org/10.3168/jds.2019-16265

Ollivier L. *Eléments de génétique quantitative: 2e édition revue et augmentée*. [s.l.] : Editions Quae, 2002. 184 p.

Ong A.-L., Teh C.-K., Mayes S., Massawe F., Appleton D. R., Kulaveerasingam H. An Improved Oil Palm Genome Assembly as a Valuable Resource for Crop Improvement and Comparative Genomics in the Arecoideae Subfamily. *Plants*. **2020**. Vol. 9, n°11, p. 1476. https://doi.org/10.3390/plants9111476

Ong-Abdullah M., Ordway J. M., Jiang N., Ooi S.-E., Kok S.-Y., Sarpan N., Azimi N., Hashim A. T., Ishak Z., Rosli S. K., Malike F. A., Bakar N. A. A., Marjuni M., Abdullah N., Yaakub Z., Amiruddin M. D., Nookiah R., Singh R., Low E.-T. L., Chan K.-L., Azizi N., Smith S. W., Bacher B., Budiman M. A., Van Brunt A., Wischmeyer C., Beil M., Hogan M., Lakey N., Lim C.-C., Arulandoo X., Wong C.-K., Choo C.-N., Wong W.-C., Kwan Y.-Y., Alwee S. S. R. S., Sambanthamurthi R., Martienssen R. A. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature*. **2015**. Vol. 525, n°7570, p. 533-537. https://doi.org/10.1038/nature15365

Ooi L. C.-L., Low E.-T. L., Abdullah M. O., Nookiah R., Ting N. C., Nagappan J., Manaf M. A., Chan K.-L., Halim M. A., Azizi N., Others. Non-tenera contamination and the economic impact of SHELL genetic testing in the Malaysian independent oil palm industry. *Front. Plant Sci.* **2016**. p. 771.

Ouyang W., Cao Z., Xiong D., Li G., Li X. Decoding the plant genome: from epigenome to 3D organization. *J. Genet. Genomics*. **2020**.https://doi.org/https://doi.org/10.1016/j.jgg.2020.06.007

Pérez P., De los Campos G. *BGLR: A Statistical Package for Whole Genome Regression and Prediction* [En ligne]. **2013**. Disponible sur : < http://R-Forge.R-project.org/projects/bglr/ >

Pérez P., De los Campos G., Crossa J., Gianola D. Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. *Plant Genome*. **2010**. Vol. 3, n°2, p. 106-116.

Pérez-Enciso M., Zingaretti L. M. A Guide on Deep Learning for Complex Trait Genomic Prediction. *Genes*. **2019**. Vol. 10, n°7,. https://doi.org/10.3390/genes10070553

Persa R., Ribeiro P. C. de O., Jarquin D. The use of high-throughput phenotyping in genomic selection context. *Crop Breed. Appl. Biotechnol.* **2021**. Vol. 21, n°spe, p. e385921S6. https://doi.org/10.1590/1984-70332021v21sa19

Piepho H. P., Möhring J., Melchinger A. E., Büchse A. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*. **2008**. Vol. 161, n°1-2, p. 209-228. https://doi.org/10.1007/s10681-007-9449-8

Pook T., Schlather M., Simianer H. MoBPS - Modular Breeding Program Simulator. *G3 Genes Genomes Genet.* **2020**. Vol. 10, n°6, p. 1915-1918. https://doi.org/10.1534/g3.120.401193

Pootakham W., Jomchai N., Ruang-areerate P., Shearman J. R., Sonthirod C., Sangsrakru D., Tragoonrung S., Tangphatsornruang S. Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics*. **2015**. Vol. 105, n°5-6, p. 288-295. https://doi.org/10.1016/j.ygeno.2015.02.002

Priyadarshan P. M. *Biology of Hevea Rubber* [En ligne]. [s.l.] : Springer International Publishing, 2017. Disponible sur : < https://books.google.fr/books?id=5nknDwAAQBAJ >ISBN : 978-3-319-54506-6.

Pszczola M., Strabel T., Mulder H. A., Calus M. P. L. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* **2012**. Vol. 95, n°1, p. 389-400. https://doi.org/10.3168/jds.2011-4338

Pujade-Renaud V. *Hevea brasiliensis : de la production de latex aux interactions biotiques*. Mémoire présenté pour l'obtention de l'Habilitation à Diriger des Recherches. [s.l.] : Université Blaise Pascal, Clermont-Ferrand, 2015.

R2D2 Consortium, Fugeray-Scarbel A., Bastien C., Dupont-Nivet M., Lemarié S. Why and How to Switch to Genomic Selection: Lessons From Plant and Animal Breeding Experience. *Front. Genet.* **2021**. Vol. 12, p. 629737. https://doi.org/10.3389/fgene.2021.629737

Rambolarimanana T., Ramamonjisoa L., Verhaegen D., Leong Pock Tsy J.-M., Jacquin L., Cao-Hamadou T.-V., Makouanzi G., Bouvet J.-M. Performance of multi-trait genomic selection for Eucalyptus robusta breeding program. *Tree Genet. Genomes*. **2018**. Vol. 14, n°5, p. 71. https://doi.org/10.1007/s11295-018-1286-5

Rastas P. Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics*. **2017**. Vol. 33, n°23, p. 3726-3732.

Ray S., Satya P. Next generation sequencing technologies for next generation plant breeding. *Front. Plant Sci.* **2014**. Vol. 5,. https://doi.org/10.3389/fpls.2014.00367 (consulté le 18 mars 2022)

Reif J. C., Gumpert F.-M., Fischer S., Melchinger A. E. Impact of Interpopulation Divergence on Additive and Dominance Variance in Hybrid Populations. *Genetics*. **2007**. Vol. 176, n°3, p. 1931-1934. https://doi.org/10.1534/genetics.107.074146

Rice B. R., Lipka A. E. Diversifying maize genomic selection models. *Mol. Breed.* **2021**. Vol. 41, n°5, p. 33. https://doi.org/10.1007/s11032-021-01221-4

Rincent R., Charpentier J.-P., Faivre-Rampant P., Paux E., Le Gouis J., Bastien C., Segura V. 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.* **2018**. Vol. 8, n°12, p. 3961-3972.

Rincent R., Kuhn E., Monod H., Oury F.-X., Rousset M., Allard V., Le Gouis J. Optimization of multi-environment trials for genomic selection based on crop models. *Theor. Appl. Genet.* **2017**. Vol. 130, n°8, p. 1735-1752. https://doi.org/10.1007/s00122-017-2922-4

Rincent R., Laloë D., Nicolas S., Altmann T., Brunel D., Revilla P., Rodriguez V. M., Moreno-Gonzales J., Melchinger A. E., Bauer E., Schön C.-C., Meyer N., Giauffret C., Bauland C., Jamin P., Laborde J., Monod H., Flament P., Charcosset A., Moreau L. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). *Genetics*. **2012**. Vol. 192, n°2, p. 2715-728. https://doi.org/10.1534/genetics.112.141473

Rival A., Levang P. *La palme des controverses : Palmier à huile et enjeux de développement*. Versailles, France : Ed. Quae, 2013. 98 p.(Essais : Quae). ISBN : 978-2-7592-2049-6.

Robertsen C. D., Hjortshøj R. L., Janss L. L. Genomic selection in cereal breeding. *Agronomy*. **2019**. Vol. 9, n°2, p. 95.

Rogers A. R. How Population Growth Affects Linkage Disequilibrium. *Genetics*. **2014**. Vol. 197, n°4, p. 1329-1341. https://doi.org/10.1534/genetics.114.166454

Rosa J. R. B. F., Mantello C. C., Garcia D., De Souza L. M., Da Silva Carla Cristina, Gazaffi R., Da Silva Cícero Casimiro, Toledo-Silva G., Cubry P., Garcia A. A. F., De Souza A. P., Le Guen V. QTL detection for growth and latex production in a full-sib rubber tree population cultivated under suboptimal

climate conditions. *BMC Plant Biol.* **2018**. Vol. 18, n°1, p. 223. https://doi.org/10.1186/s12870-018-1450-y

Russell JamesC., Fewster RachelM. Evaluation of the Linkage Disequilibrium Method for Estimating Effective Population Size. In : Thomson D, Cooch EvanG, Conroy MichaelJ (éd.). *Model. Demogr. Process. Mark. Popul.* [s.l.] : Springer US, 2009. p. 291-320. Disponible sur : < http://dx.doi.org/10.1007/978-0-387-78151-8_13 >ISBN : 978-0-387-78150-1.

Sánchez L., Toro M. A., García C. Improving the Efficiency of Artificial Selection: More Selection Pressure With Less Inbreeding. *Genetics*. **1999**. Vol. 151, n°3, p. 1103-1114.

Sanger F., Coulson A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **1975**. Vol. 94, n°3, p. 441-448. https://doi.org/10.1016/0022-2836(75)90213-2

Sanger F., Nicklen S., Coulson A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **1977**. Vol. 74, n°12, p. 5463-5467.

Dos Santos Scholz M. B., Kitzberger C. S. G., Pereira L. F. P., Davrieux F., Pot D., Charmetant P., Leroy T. Application of near infrared spectroscopy for green coffee biochemical phenotyping. *J. Infrared Spectrosc.* **2014**. Vol. 22, n°6, p. 411-421.

Sargolzaei M., Chesnais J. P., Schenkel F. S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*. **2014**. Vol. 15, n°1, p. 478. https://doi.org/10.1186/1471-2164-15-478

Schaeffer L. R. C. R. Henderson: Contributions to Predicting Genetic Merit. *J. Dairy Sci.* **1991**. Vol. 74, n°11, p. 4052-4066. https://doi.org/10.3168/jds.S0022-0302(91)78601-3

Schopp P., Müller D., Wientjes Y. C. J., Melchinger A. E. Genomic Prediction Within and Across Biparental Families: Means and Variances of Prediction Accuracy and Usefulness of Deterministic Equations. *G3 GenesGenomesGenetics*. **2017**. Vol. 7, n°11, p. 3571. https://doi.org/10.1534/g3.117.300076

Scossa F., Alseekh S., Fernie A. R. Integrating multi-omics data for crop improvement. *J. Plant Physiol.* **2021**. Vol. 257, p. 153352. https://doi.org/10.1016/j.jplph.2020.153352

Seyum E. G., Bille N. H., Abtew W. G., Munyengwa N., Bell J. M., Cros D. Genomic selection in tropical perennial crops and plantation trees: a review. *Mol. Breed.* **2022a**. Vol. 42, n°10, p. 58. https://doi.org/10.1007/s11032-022-01326-4

Seyum E. G., Bille N. H., Abtew W. G., Rastas P., Arifianto D., Domonhédo H., Cochard B., Jacob F., Riou V., Pomiès V., Lopez D., Bell J. M., Cros D. Genome properties of key oil palm (*Elaeis guineensis* Jacq.) breeding populations. *J. Appl. Genet.* **2022b**. Vol. 63, n°4, p. 633-650. https://doi.org/10.1007/s13353-022-00708-w

Da Silva A. K. V., Borges M. V. V., Batista T. S., Da Silva Junior C. A., Furuya D. E. G., Prado Osco L., Teodoro L. P. R., Baio F. H. R., Ramos A. P. M., Gonçalves W. N., Others. Predicting Eucalyptus Diameter at Breast Height and Total Height with UAV-Based Spectral Indices and Machine Learning. *Forests*. **2021**. Vol. 12, n°5, p. 582.

Singh R., Leslie Low E.-T., Ooi L. C.-L., Ong-Abdullah M., Chin T. N., Nagappan J., Nookiah R., Amiruddin M. D., Rosli R., Abdul Manaf M. A., Chan K.-L., Halim M. A., Azizi N., Lakey N., Smith S.

W., Budiman M. A., Hogan M., Bacher B., Van Brunt A., Wang C., Ordway J. M., Sambanthamurthi R., Martienssen R. A. The oil palm Shell gene controls oil yield and encodes a homologue of SEEDSTICK. *Nature*. **2013a**. Vol. 500, n°7462, p. 340-344. https://doi.org/10.1038/nature12356

Singh R., Low E.-T. L., Ooi L. C.-L., Ong-Abdullah M., Ting N.-C., Nookiah R., Ithnin M., Marjuni M., Mustaffa S., Yaakub Z., Others. Variation for heterodimerization and nuclear localization among known and novel oil palm SHELL alleles. *New Phytol.* **2020**. Vol. 226, n°2, p. 426-440.

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. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature*. **2013b**. Vol. 500, n°7462, p. 335-339.

Slatkin M. Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*. **2008**. Vol. 9, n°6, p. 477-485. https://doi.org/10.1038/nrg2361

Soh A. Breeding plans and selection methods in oil palm. In : *Symp. Sci. Oil Palm Breed. Proceedings1992Montpellier Fr.* [s.l.] : PORIM, 1999.

Soh A. C., Mayes S., Roberts J. A. *Oil Palm Breeding: Genetics and Genomics*. Boca Raton : CRC Press, 2017. 446 p.ISBN : 978-1-351-64604-8.

Solberg T. R., Sonesson A. K., Woolliams J. A., Meuwissen T. H. E. Genomic selection using different marker types and densities. *J. Anim. Sci.* **2008**. Vol. 86, n°10, p. 2447-2454. https://doi.org/10.2527/jas.2007-0010

Sørensen P., Edwards S. M., Rohde P. D. Genomic feature models. In : *10th World Congr. Genet. Appl. Livest. Prod. WCGALP*. [s.l.] : [s.n.], 2014.

Sousa I. C. De, Nascimento M., De Castro Sant'anna I., Teixeira Caixeta E., Ferreira Azevedo C., Damião Cruz C., Lopes da Silva F., Ruas Alkimim E., Campana Nascimento A. C., Vergara Lopes Serão N. Marker effects and heritability estimates using additive-dominance genomic architectures via artificial neural networks in Coffea canephora. *PLOS ONE*. **2022**. Vol. 17, n°1, p. e0262055. https://doi.org/10.1371/journal.pone.0262055

Sousa I. C. De, Nascimento M., Silva G. N., Nascimento A. C. C., Cruz C. D., Silva F. F. E, Almeida D. P. De, Pestana K. N., Azevedo C. F., Zambolim L., Caixeta E. T. Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. *Sci. Agric.* **2021**. Vol. 78, n°4, p. e20200021. https://doi.org/10.1590/1678-992x-2020-0021

De Souza Jr C. L. INTERPOPULATION GENETIC VARIANCES AND HYBRID BREEDING PROGRAMS. *Braz. J. Genet.* **1992**. Vol. 15, n°3, p. 643-656.

Souza L. M., Francisco F. R., Gonçalves P. S., Scaloppi Junior E. J., Le Guen V., Fritsche-Neto R., Souza A. P. Genomic Selection in Rubber Tree Breeding: A Comparison of Models and Methods for Managing G×E Interactions. *Front. Plant Sci.* **2019**. Vol. 10, p. 1353. https://doi.org/10.3389/fpls.2019.01353

Speed D., Balding D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **2014**. Vol. 24, p. 1550-1557. https://doi.org/10.1101/gr.169375.113

Steiger J. H. Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **1980**. Vol. 87, n°2, p. 245.

Stock J., Bennewitz J., Hinrichs D., Wellmann R. A Review of Genomic Models for the Analysis of Livestock Crossbred Data. *Front. Genet.* **2020**. Vol. 11,. https://doi.org/10.3389/fgene.2020.00568 (consulté le 17 septembre 2020)

Stuber C. W., Cockerham C. C. Gene effects and variances in hybrid populations. *Genetics*. **1966**. Vol. 54, n°6, p. 1279-1286.

Szabo Q., Bantignies F., Cavalli G. Principles of genome folding into topologically associating domains. *Sci. Adv.* **2019**. Vol. 5, n°4, p. eaaw1668. https://doi.org/10.1126/sciadv.aaw1668

Tan B., Grattapaglia D., Martins G. S., Ferreira K. Z., Sundberg B., Ingvarsson P. K. Evaluating the accuracy of genomic prediction of growth and wood traits in two *Eucalyptus* species and their F1 hybrids. *BMC Plant Biol.* **2017**. Vol. 17, n°1, p. 110. https://doi.org/10.1186/s12870-017-1059-6

Tchounke B., Sanchez L., Bell J. M., Cros D. Mate selection: a useful approach to maximize genetic gain and control inbreeding in genomic and conventional oil palm (Elaeis guineensis Jacq.) hybrid breeding. **under review**.

Technow F., Riedelsheimer C., Schrag TobiasA., Melchinger AlbrechtE. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor. Appl. Genet.* **2012**. Vol. 125, n°6, p. 1181-1194. https://doi.org/10.1007/s00122-012-1905-8

Technow F., Schrag T. A., Schipprack W., Bauer E., Simianer H., Melchinger A. E. Genome Properties and Prospects of Genomic Prediction of Hybrid Performance in a Breeding Program of Maize. *Genetics*. **2014**. Vol. 197, n°4, p. 1343. https://doi.org/10.1534/genetics.114.165860

Teh Boon Sung C., See Siang C., Sime Darby Research Sdn. Bhd., Malaysia. Modelling crop growth and yield in palm oil cultivation. In : Rival A (éd.). *Achiev. Sustain. Cultiv. Oil Palm Vol. 1 Introd. Breed. Cultiv. Tech.* [s.l.] : Burleigh Dodds Science Publishing, 2018. p. 183-228.https://doi.org/10.19103/AS.2017.0018.10 (consulté le 8 avril 2022)ISBN : 978-1-78676-104-0.

Teh C.-K., Ong A.-L., Mayes S., Massawe F., Appleton D. R. Major QTLs for Trunk Height and Correlated Agronomic Traits Provide Insights into Multiple Trait Integration in Oil Palm Breeding. *Genes*. **2020**. Vol. 11, n°7, p. 826. https://doi.org/10.3390/genes11070826

Teissier M., Larroque H., Brito L. F., Rupp R., Schenkel F. S., Robert-Granié C. 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.* **2020**. Vol. 103, n°12, p. 11559-11573. https://doi.org/10.3168/jds.2020-18662

Tester M., Langridge P. Breeding Technologies to Increase Crop Production in a Changing World. *Science*. **2010**. Vol. 327, n°5967, p. 818-822. https://doi.org/10.1126/science.1183700

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. Linkage disequilibrium vs. pedigree: Genomic selection prediction accuracy in conifer species. *PLOS ONE*. **2020**. Vol. 15, n°6, p. e0232201. https://doi.org/10.1371/journal.pone.0232201

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

Tisné S., Pomiès V., Riou V., Syahputra I., Cochard B., Denis M. Identification of Ganoderma Disease Resistance Loci Using Natural Field Infection of an Oil Palm Multiparental Population. *G3 GenesGenomesGenetics*. **2017**. Vol. 7, n°6, p. 1683. https://doi.org/10.1534/g3.117.041764

Tong H., Nikoloski Z. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *J. Plant Physiol.* **2021**. Vol. 257, p. 153354. https://doi.org/10.1016/j.jplph.2020.153354

Toro M. A., Saura M., Fernandez J., Villanueva B. Accuracy of genomic within-family selection in aquaculture breeding programmes. *J. Anim. Breed. Genet.* **2017**. Vol. 134, n°3, p. 256-263. https://doi.org/10.1111/jbg.12272

Toro M., Perez-Enciso M. Optimization of selection response under restricted inbreeding. *Genet. Sel. Evol.* **1990**. Vol. 22, n°1, p. 93-107.

Tran D. M., Clément-Demange A., Déon M., Garcia D., Le Guen V., Clément-Vidal A., Soumahoro M., Masson A., Label P., Le M. T., Pujade-Renaud V. Genetic determinism of sensitivity to *Corynespora cassiicola* exudates in rubber tree (*Hevea brasiliensis*). *PloS One*. **2016**. Vol. 11, n°10, p. e0162807-e0162807. https://doi.org/10.1371/journal.pone.0162807

Tyczewska A., Woźniak E., Gracz J., Kuczyński J., Twardowski T. Towards Food Security: Current State and Future Prospects of Agrobiotechnology. *Trends Biotechnol.* **2018**. Vol. 36, n°12, p. 1219-1229. https://doi.org/10.1016/j.tibtech.2018.07.008

USDA. *Oilseeds: world market and trade.* [En ligne]. [s.l.] : [s.n.], 2022. (Foreign Agricultural Service, Circular Series). Disponible sur : < https://apps.fas.usda.gov/psdonline/circulars/oilseeds.pdf > (consulté le 17 février 2022)

Van Eeuwijk F. A., Bustos-Korts D., Millet E. J., Boer M. P., Kruijer W., Thompson A., Malosetti M., Iwata H., Quiroz R., Kuppe C., Others. Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. *Plant Sci.* **2019**. Vol. 282, p. 23-39.

VanRaden P. M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **2008**. Vol. 91, n°11, p. 4414-4423. https://doi.org/10.3168/jds.2007-0980

VanRaden P. M. Genomic measures of relationship and inbreeding. *Interbull Bull.* **2007**. Vol. 37, p. 33-36.

Varshney R. K., Roorkiwal M., Sorrells M. E. *Genomic Selection for Crop Improvement*. 1re éd. Cham, Switzerland : Springer International Publishing, 2017. 258 p.ISBN : 978-3-319-63168-4.

Vaysse L., Bonfils F., Sainte-Beuve J., Cartault M. Natural rubber. In : *Polym. Sci. Compr. Ref. Polym. Sustain. Environ. Green Energy K Matyjaszewski M Möller Eds*. [s.l.] : Elsevier Amsterdam, 2012. p. 281-293.

Vitezica Z. G., Varona L., Elsen J.-M., Misztal I., Herring W., Legarra A. Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. *Genet. Sel. Evol.* **2016**. Vol. 48, n°1, p. 6. https://doi.org/10.1186/s12711-016-0185-1

Voss-Fels K. P., Cooper M., Hayes B. J. Accelerating crop genetic gains with genomic selection. *Theor. Appl. Genet.* **2019**. Vol. 132, n°3, p. 669-686. https://doi.org/10.1007/s00122-018-3270-8

Walsh B., Lynch M. *Evolution and selection of quantitative traits*. New York, NY : Oxford University Press, 2018. 1459 p.ISBN : 978-0-19-883087-0.

Wang G., Meng Q., Xia B., Zhang S., Lv J., Zhao D., Li Y., Wang X., Zhang L., Cooke J. P., Cao Q., Chen K. TADsplimer reveals splits and mergers of topologically associating domains for epigenetic regulation of transcription. *Genome Biol.* **2020**. Vol. 21, n°1, p. 84. https://doi.org/10.1186/s13059-020-01992-7

Wang L., Lee M., Yi Wan Z., Ye B., Alfiko Y., Rahmadsyah R., Purwantomo S., Song Z., Suwanto A., Hua Yue G. Chromosome-level Reference Genome Provides Insights into Divergence and Stress Adaptation of the African Oil Palm. *Genomics Proteomics Bioinformatics*. **2022**.https://doi.org/10.1016/j.gpb.2022.11.002

Wang X., Xu Y., Hu Z., Xu C. Genomic selection methods for crop improvement: Current status and prospects. *Crop J.* **2018**. Vol. 6, n°4, p. 330-340. https://doi.org/10.1016/j.cj.2018.03.001

Waples R. S., Do C. LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Mol. Ecol. Resour.* **2008**. Vol. 8, n°4, p. 753-756. https://doi.org/10.1111/j.1755-0998.2007.02061.x

Warren-Thomas E., Dolman P. M., Edwards D. P. Increasing Demand for Natural Rubber Necessitates a Robust Sustainability Initiative to Mitigate Impacts on Tropical Biodiversity. *Conserv. Lett.* **2015**. Vol. 8, n°4, p. 230-241. https://doi.org/10.1111/conl.12170

Wartha C. A., Lorenz A. J. Implementation of genomic selection in public-sector plant breeding programs: Current status and opportunities. *Crop Breed. Appl. Biotechnol.* **2021**. Vol. 21, n°spe, p. e394621S15. https://doi.org/10.1590/1984-70332021v21sa28

Wei M., Van der Werf J. H. J., Brascamp E. W. Relationship between purebred and crossbred parameters. *J. Anim. Breed. Genet.* **1991**. Vol. 108, n°1-6, p. 262-269. https://doi.org/10.1111/j.1439-0388.1991.tb00184.x

Weir B. S. Inferences about linkage disequilibrium. *Biometrics*. **1979**. Vol. 35, p. 235-254.

Weir B. S. *Genetic data analysis*. 2e éd. Sunderland, MA : Sinauer Associates, 1996. 445 p.

Wientjes Y. C. J., Veerkamp R. F., Calus M. P. L. The Effect of Linkage Disequilibrium and Family Relationships on the Reliability of Genomic Prediction. *Genetics*. **2013**. Vol. 193, n°2, p. 621-631. https://doi.org/10.1534/genetics.112.146290

Wiggans G. R., Cole J. B., Hubbard S. M., Sonstegard T. S. Genomic Selection in Dairy Cattle: The USDA Experience. *Annu. Rev. Anim. Biosci.* **2017**. Vol. 5, n°1, p. 309-327. https://doi.org/10.1146/annurev-animal-021815-111422

Wong C. K., Bernardo R. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor. Appl. Genet.* **2008**. Vol. 116, n°6, p. 815-824. https://doi.org/10.1007/s00122-008-0715-5

Würschum T., Longin C. F. H., Hahn V., Tucker M. R., Leiser W. L. Copy number variations of *CBF* genes at the *Fr-A2* locus are essential components of winter hardiness in wheat. *Plant J.* **2017a**. Vol. 89, n°4, p. 764-773. https://doi.org/10.1111/tpj.13424

Würschum T., Maurer H. P., Weissmann S., Hahn V., Leiser W. L. Accuracy of within- and among-family genomic prediction in triticale. *Plant Breed.* **2017b**. Vol. 136, n°2, p. 230-236. https://doi.org/10.1111/pbr.12465

Xiang T., Christensen O. F., Vitezica Z. G., Legarra A. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. *Genet. Sel. Evol.* **2016**. Vol. 48, n°1, p. 92. https://doi.org/10.1186/s12711-016-0271-4

Xu Y., Liu X., Fu J., Wang H., Wang J., Huang C., Prasanna B. M., Olsen M. S., Wang G., Zhang A. Enhancing Genetic Gain through Genomic Selection: From Livestock to Plants. *Plant Commun.* **2020**. Vol. 1, n°1, p. 100005. https://doi.org/10.1016/j.xplc.2019.100005

Yamashita S., Takahashi S. Molecular Mechanisms of Natural Rubber Biosynthesis. *Annu. Rev. Biochem.* **2020**. Vol. 89, n°1, p. 821-851. https://doi.org/10.1146/annurev-biochem-013118-111107

Yarra R., Cao H., Jin L., Mengdi Y., Zhou L. CRISPR/Cas mediated base editing: a practical approach for genome editing in oil palm. *3 Biotech*. **2020**. Vol. 10, n°7, p. 306. https://doi.org/10.1007/s13205-020-02302-5

Yeap W.-C., Norkhairunnisa Che Mohd Khan, Norfadzilah Jamalludin, Muad M. R., Appleton D. R., Harikrishna Kulaveerasingam. An Efficient Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)/CRISPR-Associated Protein 9 Mutagenesis System for Oil Palm (Elaeis guineensis). *Front. Plant Sci.* **2021**. Vol. 12, p. 773656. https://doi.org/10.3389/fpls.2021.773656

Yuan X., Miller D. J., Zhang J., Herrington D., Wang Y. An overview of population genetic data simulation. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **2012**. Vol. 19, n°1, p. 42-54. https://doi.org/10.1089/cmb.2010.0188

Yuan Y., Bayer P. E., Batley J., Edwards D. Current status of structural variation studies in plants. *Plant Biotechnol. J.* **2021**. Vol. 19, n°11, p. 2153-2163. https://doi.org/10.1111/pbi.13646

Yue G. H., Ye B. Q., Lee M. Molecular approaches for improving oil palm for oil. *Mol. Breed.* **2021**. Vol. 41, n°3, p. 22. https://doi.org/10.1007/s11032-021-01218-z

Zhao T., Zeng J., Cheng H. Extend mixed models to multilayer neural networks for genomic prediction including intermediate omics data. *Genetics*. **2022**. p. iyac034. https://doi.org/10.1093/genetics/iyac034

Zhao Y., Mette M. F., Gowda M., Longin C. F. H., Reif J. C. Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity*. **2014**. Vol. 112, n°6, p. 638-645.

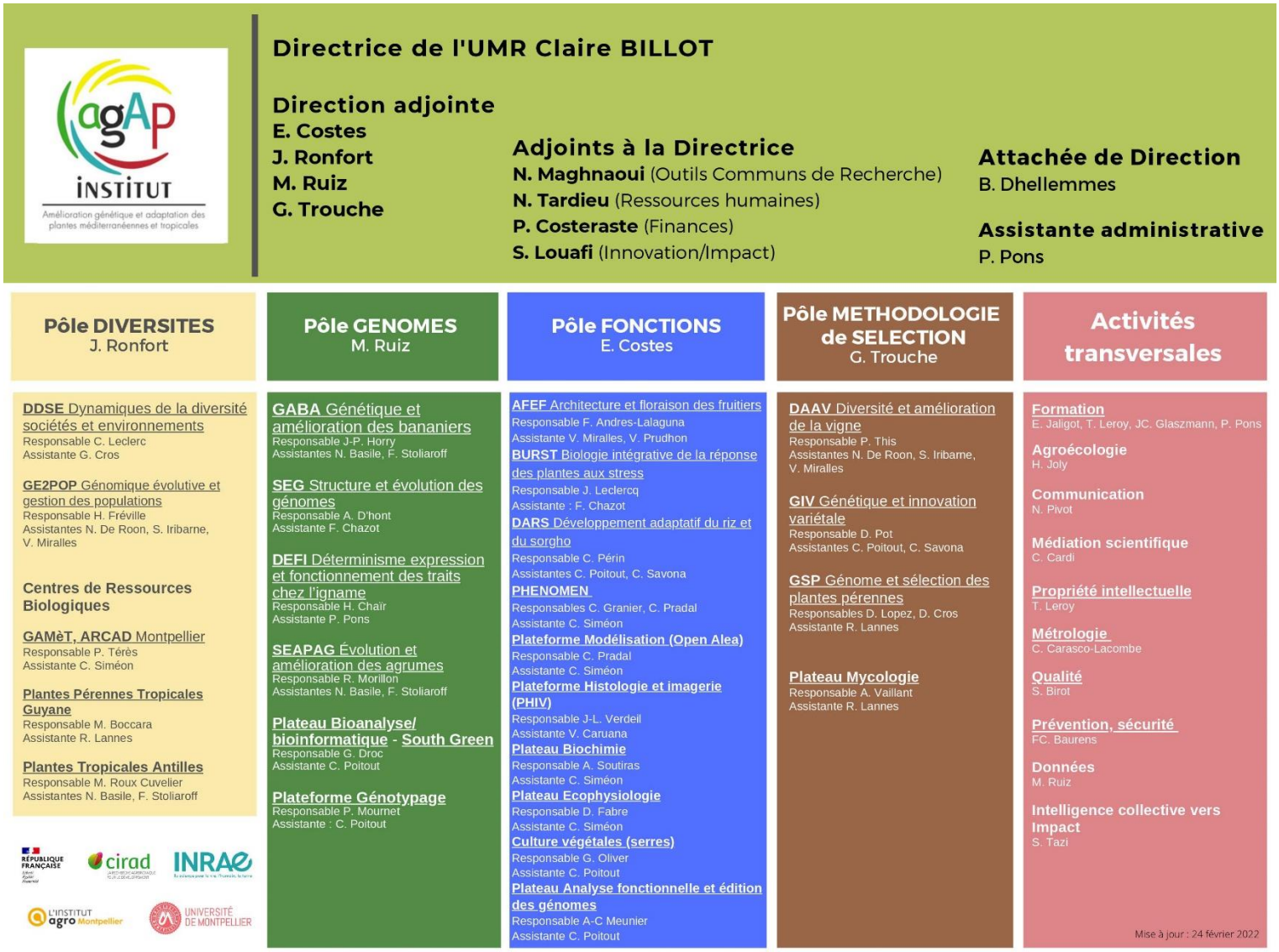
Zheng Y., Chen C., Liang Y., Sun R., Gao L., Liu T., Li D. Genome-wide association analysis of the lipid and fatty acid metabolism regulatory network in the mesocarp of oil palm (Elaeis guineensis Jacq.) based on small noncoding RNA sequencing. *Tree Physiol.* **2019**. Vol. 39, n°3, p. 356-371. https://doi.org/10.1093/treephys/tpy091

Zhou L., Holliday J. A. Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture. *BMC Genomics*. **2012**. Vol. 13, n°1, p. 703. https://doi.org/10.1186/1471-2164-13-703

Zingaretti L. M., Gezan S. A., Ferrão L. F. V., Osorio L. F., Monfort A., Muñoz P. R., Whitaker V. M., Pérez-Enciso M. Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species. *Front. Plant Sci.* **2020**. Vol. 11,. https://doi.org/10.3389/fpls.2020.00025

Annexes

Annexe 1. Organigramme de l'UMR AGAP Institut (2022)



Annexe 2. Listes des missions à l'étranger effectuées entre 2012 et mars 2022

Date début	Date fin	Pays principal	Objet
26/10/2022	07/11/2022	Cameroun	Encadrement doctorants sélection génomique palmier à huile
01/05/2022	08/05/2022	Côte d'ivoire	Démarrage de la thèse de Daouda Kouassi
25/10/2021	05/11/2021	Cameroun	Encadrement doctorants sélection génomique palmier à huile et soutenance thèse A. Nyouma
17/10/2020	27/10/2020	Cameroun	Encadrement doctorants sélection génomique palmier à huile
09/01/2020	27/01/2020	Etats-Unis	Conférence Plant & Animal Genome, San Diego, CA, USA
05/11/2019	21/11/2019	Malaisie	Projet OPGP (animation workshop sélection génomique palmier à huile), conférence PIPOC
06/04/2019	16/04/2019	Malaisie	Projet OPGP (animation workshop sélection génomique palmier à huile)
31/01/2019	01/02/2019	Bénin	Soutenance thèse Hubert Domonhedo
15/01/2019	23/01/2019	Bénin	Enseignements masters UAC et encadrement doctorat H. Domonhedo
19/04/2017	05/05/2017	Bénin	Enseignements masters UAC et encadrement doctorat H. Domonhedo
24/04/2016	12/05/2016	Bénin	Enseignement master UAC
06/03/2016	16/03/2016	Bénin	Encadrement thèse Hubert Domonhedo (CRAPP)
08/01/2016	15/01/2016	Etats-Unis	Conférence Plant and Animal Genome (San Diego)
07/03/2015	16/03/2015	Malaisie	Projet OPGP (animation workshop sélection génomique palmier à huile), conférence PIPOC
09/01/2015	16/01/2015	Etats-Unis	Conférence Plant & Animal Genome XXIII, San Diego, CA, USA
08/06/2014	20/06/2014	Indonésie	Projet OPGP (workshop), conférences IOPC et ISOPB
06/04/2014	16/04/2014	Bénin	Enseignement master UAC
09/11/2013	22/11/2013	Malaisie	Atelier projet OPGP, visite MPOB, conférence PIPOC
21/04/2013	29/04/2013	Bénin	Enseignement master UAC
11/01/2013	18/01/2013	Etats-Unis	Conférence Plant & Animal Genome XXI, San Diego, CA, USA
15/04/2012	23/04/2012	Bénin	Enseignement master UAC

Annexe 3. Modèle mixte et BLUP généalogique

Le modèle mixte est un modèle statistique reliant des observations à des effets fixes et à des effets aléatoires. Henderson (1950) a développé une méthode d'analyse des modèles mixtes donnant les solutions des effets fixes (BLUE, pour *best linear unbiased estimators*) et aléatoires (BLUP, pour *best linear unbiased predictors*). Cette méthodologie est très largement décrite dans la littérature (voir par exemple les ouvrages de Mrode (2014), Walsh et Lynch (2018) et Ollivier (2002)).

Dans les évaluations génétiques, le modèle mixte est utilisé pour prédire un vecteur de valeurs génétiques aléatoires non observables (u) à partir d'un vecteur de données (y). Les valeurs génétiques sont généralement la valeur génétique additive des individus observés ou l'aptitude générale à la combinaison de leurs parents. Le modèle mixte linéaire peut s'écrire :

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}u + e$$

avec y ($n \times 1$) le vecteur des observations, β ($p \times 1$) le vecteur des effets fixes et $\boldsymbol{X}$ ($n \times p$) sa matrice d'incidence, u ($q \times 1$) le vecteur des effets aléatoires génétiques et $\boldsymbol{Z}$ ($n \times q$) sa matrice d'incidence, et e ($n \times 1$) le vecteur des erreurs résiduelles, avec n le nombre d'observations et p et q le nombre d'effets fixes et aléatoires à estimer, respectivement.
Il suppose que u et e suivent des lois normales indépendantes :

$$\begin{bmatrix} u \\ e \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{V_u} & 0 \\ 0 & \boldsymbol{V_e} \end{bmatrix}\right)$$

avec $\boldsymbol{V_u}$ et $\boldsymbol{V_e}$ les matrices de variance-covariance génétique et résiduelle, respectivement.
Ceci implique que $y \sim N(\boldsymbol{X}\beta, \boldsymbol{V_p})$, avec une matrice de variance-covariance phénotypique $\boldsymbol{V_p} = \boldsymbol{Z}\boldsymbol{V_u}\boldsymbol{Z}^T + \boldsymbol{V_e}$. La matrice de variance-covariance associée à u est traditionnellement $\boldsymbol{V_u} = \boldsymbol{A}\sigma^2_a$, avec $\boldsymbol{A}$ la matrice d'apparentement généalogique impliquant les individus observés et leurs ascendants présents dans le pédigrée, remontant en théorie jusqu'à une population de base composée d'individus non apparentés et non sélectionnés, et σ^2_a la variance génétique additive de la population de base. La matrice de variance-covariance résiduelle ($n \times n$) est $\boldsymbol{V_e} = \boldsymbol{I}\sigma^2_e$ avec $\boldsymbol{I}$ matrice identité. Les composantes de la variance ($\boldsymbol{V_u}$ et $\boldsymbol{V_e}$) doivent être estimées avant de pouvoir obtenir les solutions pour les effets fixes (β) et aléatoires (u). Par convention, les solutions des effets fixes sont nommées estimateurs (BLUE, $\hat{\beta}$) et les solutions des effets aléatoires prédicteurs (BLUP, $\hat{u}$). Les variances sont généralement estimées par la méthode du maximum de vraisemblance restreint (REML). Elles sont ensuite utilisées pour prédire les effets aléatoires et estimer les effets fixes, grâce aux équations du modèle mixte d'Henderson (1984 ; 1986) :

$$\begin{bmatrix} \boldsymbol{X^T V_e^{-1} X} & \boldsymbol{X^T V_e^{-1} Z} \\ \boldsymbol{Z^T V_e^{-1} X} & \boldsymbol{Z^T V_e^{-1} Z + V_u^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^T V_e^{-1}} y \\ \boldsymbol{Z^T V_e^{-1}} y \end{bmatrix}.$$

Comme $\boldsymbol{V_e^{-1}}$ est une matrice identité, elle peut se mettre en facteur pour donner une forme simplifiée (Mrode, 2005, p. 41) qui s'écrit :

$$\begin{bmatrix} \boldsymbol{X^T X} & \boldsymbol{X^T Z} \\ \boldsymbol{Z^T X} & \boldsymbol{Z^T Z + A^{-1}}\lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^T} y \\ \boldsymbol{Z^T} y \end{bmatrix},$$

avec $\lambda = \sigma^2_e / \sigma^2_a$.
Enfin, les BLUE des effets fixes sont : $\hat{\beta} = (\boldsymbol{X^T \hat{V}_p^{-1} X})^{-1}\, \boldsymbol{X^T \hat{V}_p^{-1}} y$ et les BLUP des effets aléatoires sont : $\hat{u} = \boldsymbol{\hat{V}_u Z^T \hat{V}_p^{-1}}\, (y - \boldsymbol{X}\hat{\boldsymbol{\beta}})$ (Mrode, 2014, appendix C2).

Le nom du BLUP provient de ses caractéristiques :

- il maximise la corrélation entre les valeurs additives vraies (u) et prédites ($\hat{u}$) [meilleur]. Autrement dit, il minimise $\sigma^2_{(u-\hat{u})}$, la variance des erreurs de prédictions (PEV, pour *prediction error variance*),
- les solutions $\hat{\beta}$ et $\hat{u}$ sont des fonctions linéaires des observations y [linéaire],
- $E(u - \hat{u}) = 0$ [non biaisé].

La méthode du BLUP appliquée aux évaluations génétiques présente plusieurs avantages qui l'ont rendu très populaire (Piepho et al., 2008 ; Soh, 1999) :

- elle utilise la covariance additive entre tous les individus observés, quel que soit leur niveau d'apparentement (dans la mesure où il est non nul), pour améliorer la précision de l'estimation des composantes de la variance et des valeurs additives. Cette matrice rend aussi compte de l'histoire de la population : sélection, contribution inégale des individus du pédigrée, etc.,
- elle gère facilement les effets fixes,
- elle gère facilement les dispositifs expérimentaux déséquilibrés,
- elle est flexible et peut tenir compte de nombreux effets : corrélations génétiques et résiduelles entre caractères, mesures répétées, groupes génétiques, variances hétérogènes, effets maternels, interactions entre génotypes et environnement, effets de compétition, effets spatiaux, etc.

Cette méthode possède cependant certains inconvénients (Piepho et al., 2008) :

- elle suppose que les composantes de la variance soient connues sans erreur. Par conséquent, l'erreur liée à l'estimation des variances n'est pas prise en compte et elle introduit une erreur dans les solutions (notons qu'une approche Bayésienne peut régler ce problème).
- les BLUP des individus sont reserrés autour de la moyenne de leurs parents, ce qui augmente la probabilité de sélectionner des individus apparentés et abouti à un accroissement de la consanguinité plus fort qu'avec une sélection ne tenant pas compte des apparentements,
- elle repose sur plusieurs hypothèses fréquemment non vérifiées. Notamment, elle suppose que le pédigrée renvoie à une population de base idéale, composée d'individus non apparentés et non sélectionnés, et qu'il reflète donc toute l'histoire de la population. Dans la pratique, la méthode considère comme population de base les ascendants les plus lointains jusqu'où remonte le pédigrée (c-à-d sans parents connus). Des erreurs dans le pédigrée peuvent aussi biaiser les résultats. Par ailleurs, le BLUP suppose que les termes de ségrégation mendélienne sont aléatoires, ce qui signifie notamment qu'il ne doit pas y avoir eu de sélection des individus ; sauf si toutes les informations utilisées pour prendre les décisions de sélection sont disponibles (Cantet et Fernando, 1995). Cependant, le BLUP a montré une grande robustesse face à la violation de ces hypothèses, ce qui a contribué à son succès.

Annexe 4. Palmier à huile (*Elaeis guineensis* Jacq) et production d'huile

A

http://commons.wikimedia.org

B

http://www.nafas.com.my

C

A Plantation commerciale (Malaisie)
B Récolte (Malaisie)
C Récolte (Bénin)
D Inflorescence mâle à maturité
E Régime mûr
F Fruits mûrs

D

E

F

G

H

http://commons.wikimedia.org

I

G Coupe transversale de fruit tenera (type commercial)
H Usine et régimes (Côte d'Ivoire)
I Plantation et usine (Indonésie)
J Presse semi-industrielle
K Presse artisanale
L Huile rouge (huile de palme brute, non raffinée)

J

K

L

Annexe 5. Amélioration génétique et production de semences chez le palmier à huile

M

M Palmier à huile élite de la population Deli (PO3600D, Pobè, Bénin)
N Inflorescence mâle dégagée avant ensachage en prévision de la récolte du pollen
O Inflorescence femelle ensachée, à maturité pour la fécondation artificielle
P Fécondation artificielle
Q Couronne de palmier à huile de la population Deli chargée de régimes de fécondations artificielles
R Régime de fécondation artificielle proche de la maturité (Deli) et sélectionneur (Bruno Nouy)
S Graines sèches
T Graines germées
U Coupe longitudinale de fruit dura
V Coupe transversale de fruit pisifera
W Composantes du régime (pédoncule, épillets, fruits et graines)
X Décompte et pesée des régimes
Y Mesure de hauteur
Z Extracteurs de Soxhlet destinés à mesurer le pourcentage d'huile dans la pulpe

N

O

P

Q

R

S

T

U

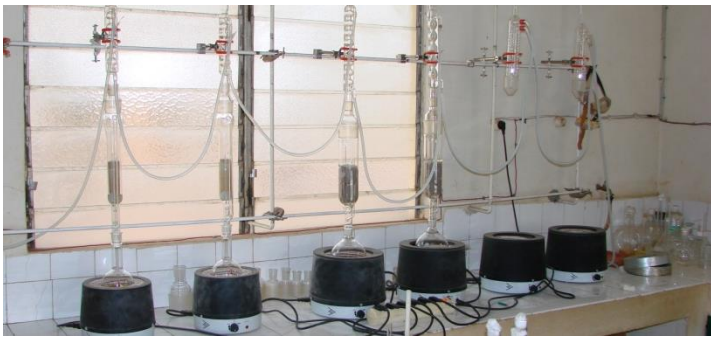
V

W

X

Y

Z



Annexe 6. Hévéa (*Hevea brasiliensis* Müll. Arg.), production de latex et amélioration génétique

A

B

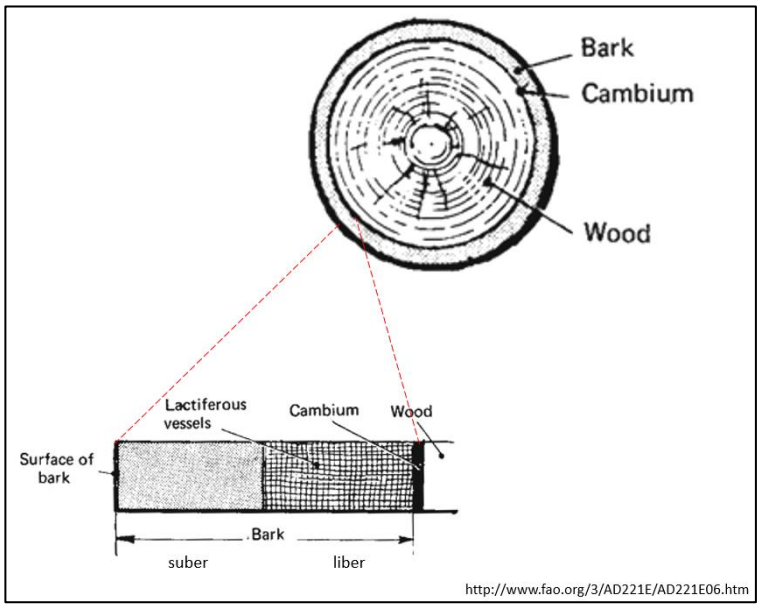
https://powo.science.kew.org © RBG Kew

© Giuseppe Mazza

C

D

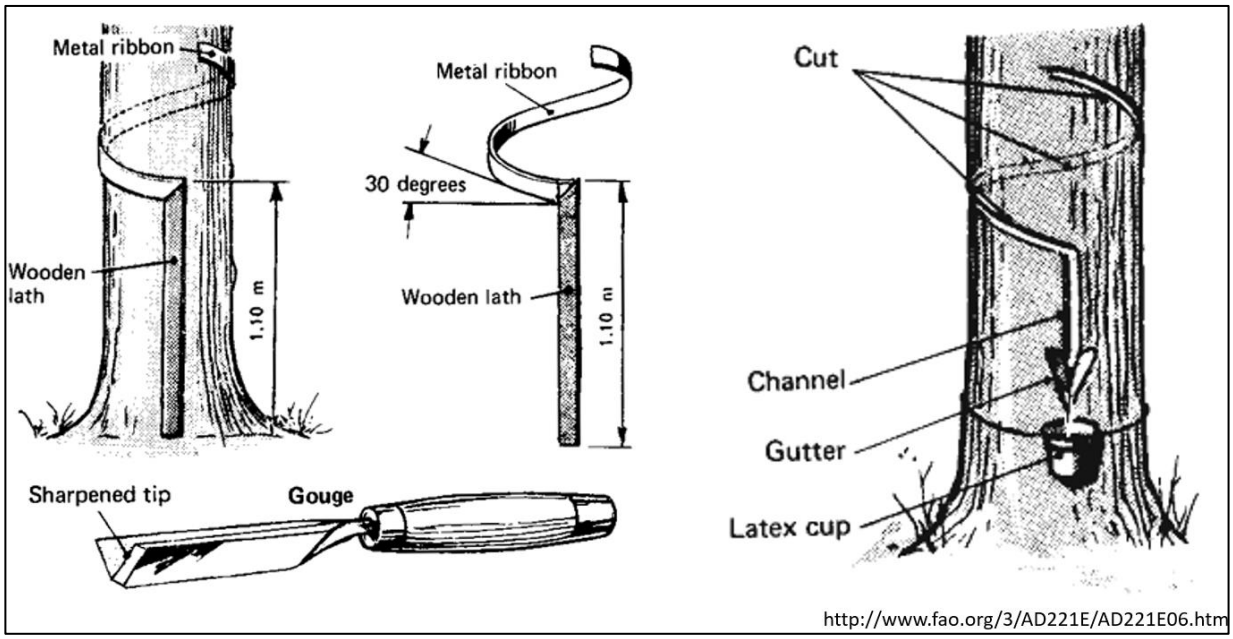
E



A Plantation d'hévéas greffé (clone GT1)
B haut : Inflorescence,
bas : Fleurs
C Fruits
D Graines
E Structure du tronc et localisation des vaisseaux laticifères

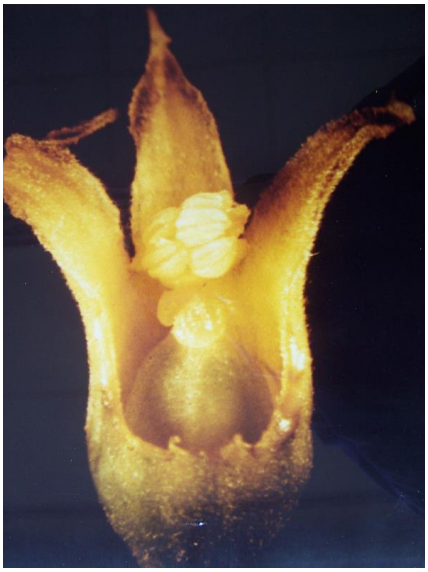
(continue)

F



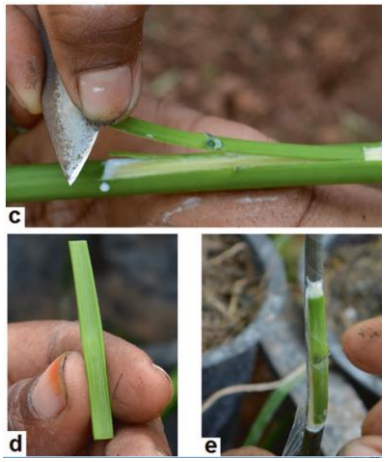
G

H



F Schéma de principe de la saignée
G Arbre en saignée
H Coupe d'une fleur en fécondation artificielle, avec des anthères mises en contact avec les stigmates
I Greffe de bourgeon
J Greffon sur porte-greffe
K Pesée de la production de latex

I



J

K

(Priyadarshan, 2017)

Annexe 7. Le déséquilibre de liaison

Plusieurs méthodes permettent de mesurer le DL (Russell et Fewster, 2009 ; Slatkin, 2008 ; Weir, 1979 ; Weir, 1996). Considérons deux loci avec le premier possédant un allèle *A* et le second un allèle *B* (quel que soit le nombre d'allèles par locus). Le DL se rattache à la grandeur D_{AB}, qui est la déviation entre la fréquence observée de l'haplotype *AB* et sa fréquence attendue sous l'hypothèse d'indépendance des allèles *A* et *B* et de reproduction au hasard. Si la transmission des allèles est indépendante entre les deux loci, c'est-à-dire en équilibre de liaison, alors l'haplotype *AB* doit se rencontrer à une fréquence (p_{AB}) égale au produit de la fréquence des deux allèles impliqués ($p_A p_B$). Le DL entre les allèles *A* et *B* est l'écart entre la fréquence observée et la fréquence attendue :

$$D_{AB} = p_{AB} - p_A p_B$$

D_{AB} peut donc s'obtenir facilement à partir des fréquences haplotypiques. On note que pour des loci biallèliques $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$.

Dans la pratique ce sont souvent des données génotypiques qui sont disponibles. Dans ce cas, on utilise généralement Δ_{AB} (delta de Burrow) au lieu de D_{AB}. Δ_{AB} peut se calculer à partir de données génotypiques et sans faire l'hypothèse de reproduction au hasard. Dans notre exemple avec deux loci biallèliques, on a (Weir, 1979) :

$$\Delta_{AB} = \frac{n_{AB}}{n} - 2p_A p_B$$

avec n_{AB} un décompte des différents génotypes possibles (voir Russell et Fewster, 2009, p. 297), *n* le nombre d'individus échantillonnés et p_A et p_B la proportion d'allèles *A* et *B* dans l'échantillon. Un estimateur non biaisé de Δ_{AB} s'obtient en tenant compte de la taille *n* de l'échantillon : $\hat{\Delta}_{AB} = \Delta_{AB} \frac{n}{n-1}$ (Lynch et Walsh, 1998, p. 98 ; Weir, 1979). Comme D_{AB} et $\hat{\Delta}_{AB}$ sont sensibles aux fréquences allèliques, il est difficile de les utiliser pour comparer le DL entre différentes paires de loci. On les transforme donc en une grandeur standardisée (*r*) qui correspond au coefficient de corrélation entre les allèles des loci concernés. Par ailleurs comme *r* peut être positif ou négatif, on lui préfère la grandeur r^2 lorsque l'on s'intéresse à l'amplitude du DL. Dans notre exemple, on a (Slatkin, 2008 ; Weir, 1996) :

$$r^2_{D_{AB}} = D^2_{AB} \,/\, (p_A(1-p_A)p_B(1-p_B))$$

et (Weir, 1996) :

$$r^2_{\hat{\Delta}_{AB}} = \hat{\Delta}_{AB} \,/\sqrt{(p_A(1-p_A) + (p_{AA} - p_A^2))(p_B(1-p_B) + (p_{BB} - p_B^2))}\,.$$

De nombreux facteurs influencent le DL (Flint-Garcia et al., 2003 ; Gupta et al., 2005 ; Hamilton, 2021 ; Mackay et Powell, 2007 ; Rogers, 2014 ; Slatkin, 2008). Les facteurs augmentant le DL sont liés à la dérive, à la constitution de la population et à la sélection, naturelle ou artificielle : consanguinité, petit nombre d'individus (dérive génétique aléatoire), isolement reproductif entre groupes d'individus, mélange de populations (*admixture*), goulots d'étranglement (réduction extrême dans la taille de la population), etc. Le DL se réduit au fil des générations par les recombinaisons, qui rompent les haplotypes existants. Parmi les facteurs accélérant le rythme de réduction du DL, on trouve les régimes de reproduction privilégiant les individus non apparentés, un taux élevé de mutation, etc. D'autres facteurs peuvent augmenter ou diminuer le DL, selon les situations, ou peuvent augmenter le DL entre certains loci et le diminuer pour d'autres. Par exemple, les mutations peuvent aboutir à un DL élevé entre allèles mutants et à un DL faible entre allèles mutants et sauvages. Enfin, certains facteurs peuvent avoir un effet uniquement sur le DL de régions particulières du génome. Par exemple, les mutations favorables créent du DL au niveau local car les allèles mutants sélectionnés et des allèles neutres physiquement proches se répandent conjointement dans la population (autostop génétique ou *genetic hitchhiking*).

L'étendue du DL correspond à la distance physique (en paires de bases) ou génétique (en Morgan) en dessous de laquelle le DL est considéré comme significatif (par exemple supérieur à 0.1 lorsque l'on mesure le DL par le r^2 entre paires de marqueurs adjacents).