Sujet de thèse

Titre : **Extension des possibilités de la sélection génomique chez le palmier à huile par l'intégration de données moléculaires individuelles d'hybrides**

Période : sept 2017 – sept 2020

Ecole doctorale : Unité de recherche et de formation doctorale (URFD) mathématique, informatique, bioinformatique et applications, Université Yaoundé 1

Contexte et objectifs :

Le schéma actuel d'amélioration du palmier à huile a permis un progrès génétique considérable sur le rendement depuis les années 1950 (Durand-Gasselin et al., 2010). Il passe par l'estimation de la valeur génétique additive d'individus des groupes A et B candidats à la sélection, obtenue grâce à des tests sur descendance dans lesquels ces candidats sont croisés pour produire des descendants hybrides A x B observés en essais. Une analyse par un modèle linéaire mixte permet alors d'estimer les valeurs additives. Les croisements hybrides commercialisés seront par la suite produits à partir des individus A et B ayant les meilleures valeurs additives. Cette méthode met donc essentiellement à profit la variabilité génétique additive au sein des populations A et B. Bien qu'il s'agisse de la principale source de variabilité génétique, deux autres sources pourraient être (mieux) exploitées pour accroitre encore le progrès génétique :

- la variabilité génétique (additive et non additive) intra-croisements hybrides, qui pourrait être valorisée en clonant les meilleurs individus hybrides. Ceci a longtemps été freiné par l'anomalie *mantled* liée au clonage, mais de récentes avancées devraient faire disparaitre rapidement ce problème (Ong-Abdullah et al 2015), relançant l'intérêt de cette stratégie. Une méthode permettant d'estimer avec précision la valeur génétique totale des individus hybrides sera donc nécessaire, afin de choisir efficacement ceux à utiliser comme têtes de clones. Actuellement, ceci se fait sur la base de leur valeur propre, une méthode peu précise compte tenu de la faible héritabilité de plusieurs composantes du rendement (effet non négligeable du micro-environnement, qui vient masquer la valeur génétique). Une autre méthode consiste à cloner les individus hybrides candidats têtes de clones, ce qui permet d'obtenir une estimation précise de leur valeur génétique mais qui rallonge d'une dizaine d'année le processus. Aujourd'hui, on peut espérer que les approches de sélection génomique (SG) (Meuwissen et al., 2001), c-à-d utilisant des données moléculaires dans le modèle mixte, donnent des estimations de la valeur génétique d'individus hybrides de manière précise et rapide, permettant de mieux identifier les têtes de clones à sélectionner pour une sortie variétale. Ceci devrait être possible en utilisant un modèle de SG incluant des données génomiques des hybrides candidats têtes de clones. Dans une première partie, cette thèse prévoit d'implémenter et de tester empiriquement l'efficacité d'un tel modèle.

- la variabilité génétique non additive entre croisements, qui est souvent négligée car les estimations des effets non additifs sont moins précises que pour les effets additifs.

En effet, compte tenu de la lourdeur des tests sur descendance, seule une très petite proportion des croisements A x B possibles est observée en essais. Pour les autres, compte tenu de l'absence d'observations, la précision des estimations des effets non additifs est très faible. Marchal et al (2016) ont montré que la théorie indiquait que cette précision serait augmentée par un modèle de SG incluant des données moléculaires des parents A et B. Par ailleurs, des simulations ont montré qu'utiliser aussi des données moléculaires sur les individus hybrides améliorait la précision des valeurs additives (Cros et al, 2016). Cela laisse supposer que des données moléculaires sur les hybrides rendraient aussi plus précises les estimations des effets non additifs. Dans une seconde partie, cette thèse prévoit de vérifier empiriquement si la précision des valeurs additives des parents A et B est augmentée par l'utilisation de données moléculaires d'hybrides ; ainsi que d'étudier si ces données améliorent aussi la précision des valeurs génétiques non additives des croisements.

En résumé, le but de cette thèse sera d'évaluer l'intérêt pour l'approche génomique appliquée au palmier à huile d'utiliser des données génomiques d'individus hybrides A x B. Deux utilisations de telles données seront étudiées :

- prédire la valeur génétique des individus hybrides A x B, pour sélectionner des têtes de clones,
- améliorer l'estimation des valeurs génétiques additives des parents A et B et des valeurs génétiques non additives des croisements hybrides A x B.

Dans les deux cas, l'effet de nombreux paramètres sera à étudier : modèle (en particulier modélisation des effets non additifs), méthode statistique d'analyse, densité de marquage moléculaire, choix des marqueurs, méthode de phasage des données moléculaires, caractère, etc.

Détail du travail prévu :

1ère partie : Utilisation de données génomiques d'individus hybrides A x B pour améliorer la sélection des têtes de clones

Ce travail sera réalisé grâce aux données concernant un ensemble de 42 clones de la société PalmElit. Ces clones ont été obtenus à partir de 42 individus hybrides (utilisés comme têtes de clones), et ont été observés en essais. On utilisera les données phénotypiques individuelles des têtes de clones, les données phénotypiques de leurs clones en essais et leurs données moléculaires (SNP obtenus par génotypage-par-séquençage). Toutes ces données sont disponibles.

Le principal enjeu est de bien estimer la valeur génétique totale des têtes de clones, afin de garantir que les performances des clones qui en seront tirés seront effectivement supérieures à celles des croisements classiques (obtenus par reproduction sexuée). Dans travail, on comparera quatre méthodes pour estimer la valeur génétique totale des têtes de clones, dont deux nouvelles méthodes permises par les modèles génomiques. On considérera les 42 clones de l'étude comme

des candidats têtes de clones, pour lesquels il s'agit d'identifier la méthode permettant d'obtenir le meilleur compromis entre précision de l'estimation de leur valeur génétique (r), intensité de sélection (i) et nombre d'années nécessaires pour faire cette estimation (L) (= maximisation du ratio $r \times i / L$) :

- Méthode conventionnelle 1 : Les candidats têtes de clones ont été clonés et installés en essais selon des dispositifs expérimentaux permettant une analyse statistique rigoureuse, avec plusieurs répétitions par clone, chacune comportant plusieurs palmiers. Des données de production individuelles ont été collectées sur chacun de ces palmiers pendant plusieurs années. L'inconvénient de cette méthode est le temps nécessaire (clonage, installation des essais au champ puis collecte des données). L'avantage est que les données permettent d'estimer de manière très précise la valeur génétique totale de chacun des 42 clones. On considérera pour l'étude que cette méthode a une précision de 1, soit la meilleure possible, et qu'elle donne les valeurs génétiques de référence (« vraies »).

- Méthode conventionnelle 2 : La valeur génétique totale des candidats têtes de clones sera estimée sur la base de leurs données phénotypiques individuelles, et éventuellement en utilisant le phénotype d'individus hybrides apparentés (d'après la généalogie). Cette méthode est beaucoup plus rapide que la méthode précédente. Par contre, sa précision peut-être faible, en particulier pour les caractères faiblement héritables.

- Méthode alternative génomique 1 : La valeur génétique totale des candidats têtes de clones sera prédite à partir de leurs propres données (phénotypes et génotypes sur un grand nombre de marqueurs SNP) et des données d'individus qui sont les parents d'autres hybrides testés en essais (dispositifs expérimentaux « Aek Loba » et « Aek Kwasan 2 » de PalmElit). En termes de temps nécessaire à l'obtention des valeurs génétiques estimées, cette méthode est équivalente à la méthode conventionnelle 2, mais on fait l'hypothèse qu'elle sera plus précise.

- Méthode alternative génomique 2 : Cette méthode est dérivée de la précédente, mais étudie la possibilité de sélectionner des têtes de clones parmi des individus qui n'ont pas été observés en essais. Cette approche est intéressante car elle permet aussi d'augmenter l'intensité de sélection. En effet, dans les 3 méthodes précédentes les têtes de clones sont recherchées parmi les croisements A x B qui ont été installés pour évaluation en essais. Très probablement, les croisements impliquant les meilleurs parents A et B n'en font pas partie, et cela limite l'intensité de sélection des têtes de clones. A l'issue de l'analyse des essais au champ, il est par contre possible de réaliser ces croisements et d'en génotyper les plantules, dont on prédira la valeur génétique totale. La « méthode alternative génomique 2 » vise donc à identifier les têtes de clones dans une pépinière constituée d'individus des meilleurs croisements possibles. Pour cela, on prédira les valeurs génétiques totales des candidats têtes de clones de la même manière qu'avec la méthode génomique 1, mais sans utiliser leurs phénotypes.

La précision de la méthode conventionnelle 2 et des deux méthodes génomiques sera la corrélation entre les valeurs génétiques totales estimées et les valeurs vraies, fournies par la méthode conventionnelle 1. On conclura quant à la meilleure méthode en particulier sur la base de leur ratio $r \times i / L$.

On fera les estimations des valeurs génétiques totales en étudiant l'effet du modèle (en particulier concernant la modélisation des effets non additifs), de la méthode de phasage des données moléculaires hybrides, de la méthode statistique d'analyse du modèle (BLUP, approches bayésiennes, etc), du caractère, de la densité de marquage, de la méthode de sélection des marqueurs moléculaires à utiliser dans le modèle et de la population d'apprentissage (Aek Loba ou Aek Loba + Aek Kwasan 2).

Attendu : un article

2ème partie : Utilisation de données génomiques d'individus hybrides A x B pour améliorer l'estimation des valeurs additives des parents A et B et des effets non additifs des croisements hybrides A x B

Ce travail sera réalisé sur deux dispositifs expérimentaux comprenant plusieurs centaines de croisements (données fournies par la société PalmElit), génotypés avec des SNP par la méthode de génotypage-par-séquençage (GBS). Pour cette partie, des données moléculaires sont en cours d'acquisition (génotypes GBS de 470 individus hybrides A x B), en complément de celles déjà disponibles (données GBS des parents A et B).

On utilisera le premier dispositif expérimental (« Aek Loba ») pour prédire la valeur génétique additive des parents et la valeur génétique totale des croisements du second dispositif (« Aek Kwasan 2 »), en utilisant des modèles génomiques incluant des données moléculaires sur les hybrides. Par simulations (Cros et al, 2015), cette approche est apparue comme la meilleure en termes de progrès génétique annuel chez le palmier à huile. Cependant, elle n'a jamais été étudiée empiriquement. Par ailleurs, les simulations ne considéraient que trois caractères (nombre de régimes, poids moyen et production totale), alors que cette partie de la thèse portera sur toutes les composantes du rendement. Les simulations supposaient aussi un déterminisme génétique purement additif, alors qu'une part d'effets non additifs (de dominance au moins) est présente, en proportions variables selon les caractères. Enfin, toujours d'après les simulations, lorsque le nombre d'individus hybrides génotypés était réduit (300), les valeurs additives étaient moins précises que celles obtenues sans les génotypes d'hybrides ; et il fallait augmenter le nombre d'hybrides génotypés pour que cette approche devienne la meilleure. Ici, on étudiera, en faisant varier le nombre d'individus hybrides génotypés, l'origine de ce résultat contre-intuitif, a priori à rechercher dans le manière d'intégrer dans le modèle des données de parenté de natures différentes (génomiques haplotypiques chez les individus hybrides génotypés, généalogiques chez les autres individus hybrides et génomiques chez les parents) ; en essayant de trouver une méthode efficace à qui profite les données moléculaires d'hybrides génotypés, quel qu'en soit le nombre.

On se basera sur les résultats de la première partie pour le choix de la méthode de phasage, la densité de marquage et le choix des marqueurs. Il faudra par contre identifier un modèle et une

méthode statistique d'analyse adaptés, qui pourraient différer de ce qui aura été identifié dans la première partie (ici il faudra associer des données moléculaires d'hybrides génotypés et des données généalogiques d'hybrides non génotypés). On étudiera aussi l'effet du nombre d'individus hybrides génotypés et du caractère.

Attendu : un article

NB : le troisième article sera un article de revue qui fera la synthèse des études de sélection génomique appliquée au palmier à huile

Références bibliographiques :

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.

Marchal A., Legarra A., Tisné S., Carasco-Lacombe C. *et al.*, 2016. **Multivariate genomic model improves analysis of oil palm (*Elaeis guineensis* Jacq.) progeny tests**. Molecular Breeding, 36(2).

Meuwissen T.H.E., Hayes B.J. et Goddard M.E., 2001. **Prediction of total genetic value using genome-wide dense marker maps**. Genetics, 157(4): 1819-1829.