

Prix de mémoire de fin d'études de la Fondation Xavier Bernard

ANNÉE : 2025

TITRE : Analyse comparative de modèles d'association génétique pour l'intégration de données omiques dans un contexte multi-populations

AUTEUR : Vincent Spinelli

ÉCOLE : ENSAT, Toulouse

Laboratoire d'accueil : MIAT, INRAE Toulouse

Rapporteur : Alain Charcosset, Claire Rogel-Gaillard

Résumé français

Les études d'association pangénomiques (eGWAS) permettent d'identifier des variants génétiques pouvant réguler l'expression génique. Leur interprétation devient toutefois complexe lorsque les données regroupent plusieurs races différenciées et que les effets d'un même variant (eQTL) peuvent différer entre elles. Ce travail évalue la capacité des modèles eGWAS globaux et intra-races à détecter des eQTLs à effets variables entre races. À partir de génotypes réels de deux races de porcs issus du projet GENE-SWitCH, un cadre de simulation a été développé pour tester différents scénarios d'effets. Les résultats montrent que les modèles inter-races bénéficient d'une meilleure puissance lorsque les effets des variants sont identiques, mais deviennent fortement limités lorsque les effets sont différents ou opposés. Les approches intra-races détectent tous les types d'effet, au prix d'une perte de puissance. Ces résultats appellent modèles plus flexibles, notamment hiérarchiques bayésiens.

English abstract

Expression genome-wide association studies (eGWAS) can identify genetic variants that may regulate gene expression. Their interpretation becomes challenging when datasets combine multiple genetically differentiated breeds and when the effects of a given variant may differ between them. This work evaluates the ability of global and within-breed eGWAS models to detect eQTLs whose effects vary between breeds. Using real genotypes from two pig breeds from the GENE-SWitCH project, a simulation framework was developed to test different realistic effect scenarios. Results show that inter-breed models benefit from greater power when variant effects are identical across breeds, but show strongly reduced power when effects differ or are opposite. Within-breed models correctly detect all types of effects, at the cost of reduced statistical power. These findings support the exploration of more flexible approaches, including hierarchical Bayesian models.

Contexte et enjeux

La sélection intensive pratiquée chez les espèces d'élevage a permis des gains de production importants, au prix d'une érosion de la diversité génétique exploitée, réduisant les leviers d'adaptation face aux nouveaux enjeux environnementaux. Dans ce contexte, il apparaît pertinent de caractériser et d'explorer la diversité génétique négligée, potentiellement porteuse de traits d'intérêt. L'essor du séquençage à haut débit permet aujourd'hui d'étudier conjointement la diversité génétique et l'expression des gènes. Les eGWAS constituent dans ce contexte un outil prometteur pour identifier les variants impliqués dans la régulation génique, mais leur application à plusieurs races reste complexe : la structure génétique inter-races peut biaiser les analyses, et un même variant peut avoir des effets différents, voire opposés entre les races. Il est donc nécessaire d'évaluer la capacité de ces modèles à identifier correctement les variants d'intérêt, au risque de les caractériser de manière erronée, voire de ne pas les détecter.

Objectifs, méthodes et points forts

L'objectif du stage était de comparer la capacité de modèles globaux et intra-races à détecter des eQTLs selon différents scénarios d'effets entre races. Pour cela, j'ai réalisé une étude de simulation visant à générer 100 expressions géniques réalistes à partir de génotypes réels de porcs et de 5 000 variants présentant différents scénarios d'effets : identiques, différents ou opposés entre races. La réalisation de ce cadre de simulation a constitué le point fort de ce stage, de par sa flexibilité totale et sa puissance : L'effet réel des variants impactant chaque gène étant contrôlé et connu en amont, il est possible d'évaluer les résultats des modèles en distinguant les vrais positifs des faux positifs, mais aussi d'identifier les variants non détectés et d'en analyser les causes.

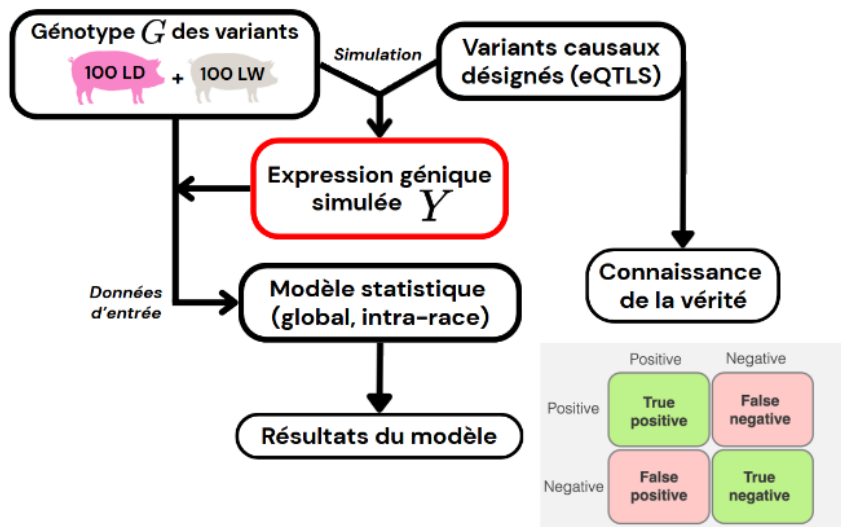


Schéma simplifié de l'étude de simulation réalisée durant le stage.

Grâce à la connaissance en amont des eQTLs simulés, chaque association (gène, variant) identifiée par les modèles peut être évaluée à l'aide d'une matrice de confusion, distinguant les vrais positifs (VP), faux positifs (FP), vrais négatifs (VN) et faux négatifs (FN).

Productions scientifiques

Développement d'un pipeline de simulation en R, modulaire et répliquable, permettant de tester différents modèles eGWAS dans un cadre contrôlé. Production d'un rapport de stage servant de base à la rédaction d'un preprint par le doctorant ayant repris le projet : <https://hal.science/hal-05651322>.

Perspectives envisagées

Les résultats ont montré le besoin de modèles capables de prendre en compte explicitement l'hétérogénéité des effets entre races. Une perspective majeure est l'étude de modèles hiérarchiques bayésiens, notamment *mashr*, qui permettent de modéliser conjointement des effets partagés, spécifiques ou corrélés entre races. L'amélioration de la simulation, notamment pour mieux représenter la variabilité génotypique et le déséquilibre de liaison, constitue une autre piste. Les travaux ont été poursuivis dans le cadre d'une thèse sur le sujet suivant :
« intégration de données multi-omiques à grande échelle. »

Votre appréciation personnelle

Ce stage a largement testé ma capacité à émettre des hypothèses et à résoudre des problèmes, dans le cadre d'un sujet particulièrement riche qui s'est développé au fil des résultats obtenus. J'en retiens notamment une conviction forte quant à la puissance des études de simulation qui constituent selon moi un environnement particulièrement pertinent pour tester des hypothèses dans des situations où la vérité est connue et pour identifier précisément les limites de méthodes utilisées.