SPÉCIFICITÉ DES ASSOCIATIONS GÉNÉTIQUES EN CONTEXTE MULTI-POPULATIONS : UNE ÉTUDE PAR SIMULATION

Kossi Julien Kowou[1,2], Vincent Spinelli[1], Andrea Rau[2], Nathalie Vialaneix[1]

[1] *Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France*
[2] *Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350, Jouy-en-Josas, France*
{kossi-julien.kowou, andrea.rau, nathalie.vialaneix}@inrae.fr,
vincentspinelli1@gmail.com

Résumé. Les études d'association à l'échelle du génome (GWAS) sont cruciales pour comprendre les liens entre variations génétiques et phénotypes complexes d'intérêt. En agriculture, la sélection a induit la structuration des espèces agricoles en sous-populations (races ou écotypes) hautement spécialisées. À ce jour, peu d'études ont évalué l'impact conjoint de cette structuration et du choix du modèle sur les résultats d'un GWAS. Dans cette étude, nous proposons un cadre de simulation réaliste dans le but d'évaluer la performance des méthodes existantes pour détecter des associations spécifiques à une population ou capturer des effets partagés ou non entre sous-populations. Les résultats obtenus dans cette étude mettent en évidence les limites des approches classiques fondées sur des modèles linéaires mixtes, notamment pour la détection de variants présentant des effets hétérogènes ou opposés entre populations.

Mots-clés. multi-population, association génétique, modèle linéaire mixte, simulation, modèle hiérarchique bayésien

Abstract. Genome-wide association studies (GWAS) are crucial for understanding the links between genetic variations and complex phenotypes of interest. In agriculture, selection has led to the structuring of agricultural species into highly specialized subpopulations (breeds or ecotypes). To date, few studies have evaluated the combined impact of this structuring and model choice on GWAS results. In this study, we propose a realistic simulation framework to evaluate the performance of existing methods for detecting population-specific associations or capturing shared or non-shared effects between subpopulations. The results obtained in this study highlight the limitations of classical approaches based on linear mixed models, particularly for the detection of variants with heterogeneous or opposing effects between subpopulations.

Keywords. multi-population, genetic association, linear mixed model, simulation, bayesian hierarchical model

1 Introduction

Les études d'associations génétiques à l'échelle du génome (GWAS : *Genome-wide association studies*) sont utilisées pour étudier la relation entre variants génétiques et phénotypes d'intérêt. Dans les espèces agricoles, la sélection a induit une structuration en races pour les animaux d'élevage et en écotypes pour les espèces végétales, que nous désignerons par la suite sous le terme générique de « populations ». Cependant, très peu d'études ont évalué l'impact conjoint de la structuration en population et du choix du modèle sur les résultats des analyses GWAS.

Une première étude de simulation a été menée par van den Berg and MacLeod (2023) pour évaluer l'influence des différences de fréquence allélique (VAF : *Variant allele frequency*) des loci de caractères quantitatifs (*Quantitative trait loci* ou QTL, des régions/points précis dans le génome associé·e·s à la variation quantitative d'un caractère phénotypique) sur la puissance et la précision des modèles d'associations utilisés. Les QTLs ont été simulés selon trois scénarios : un premier où la VAF était similaire dans les trois races bovines étudiées, un second où la VAF était variable dans les trois races,

et un troisième où les QTLs étaient spécifiques aux races. Les auteurs ont examiné l'effet de ces trois scénarios en utilisant plusieurs modèles GWAS, intra-races ou global sur toutes les races. Les résultats révèlent une insuffisance de puissance des analyses intra-races ou globalement des analyses portant sur les variants à faible VAF. En revanche, les modèles globaux ont un défaut de puissance sur les variants spécifiques d'une population. Toutefois, cette étude est insuffisante car elle ne prend pas en compte le scénario où l'association pourrait être spécifique à une population donnée.

Une autre étude préliminaire a été réalisée dans Ko et al. (2024) sur des données réelles. Le résultat principal de l'étude met en évidence des différences marquées entre les modèles inter et intra-races concernant les variants identifiés, ainsi que l'existence d'associations spécifiques ou présentant des effets de signe inverse entre races dans les données réelles. Toutefois, une limite importante réside dans le fait que l'étude ne reposait pas sur un cadre où la vérité terrain était connue, ce qui n'a pas permis d'évaluer la qualité des prédictions.

Notre objectif ici est de proposer un cadre de simulation pour pouvoir valider la performance des méthodes proposées dans la littérature, comme dans van den Berg and MacLeod (2023), tout en considérant un scénario plus riche intégrant des associations spécifiques aux populations.

2 Évaluation de la qualité des méthodes existantes

2.1 Modèles étudiés

Dans cette section, nous présentons les modèles étudiés. Dans la suite, nous noterons $(y_i)_{i=1,\dots,n}$ l'observation d'un phénotype quantitatif (par exemple, l'expression d'un gène) pour les individus $i \in \{1,\dots,n\}$ et $X = (X_{ij})_{i=1,\dots,n, j=1,\dots,p}$ la matrice des observations de p SNPs sur ces mêmes individus. Nous supposerons, en outre, que les n individus sont structurés en R sous-populations. Le vecteur $\mathbf{z} = (z_i)_{i=1,\dots,n}$ tel que $z_i \in \{1,\dots,R\}$.

Modèle linéaire mixte. Les modèles linéaires mixtes (MLM) sont largement utilisés pour tenir compte de la structure de la population dans les études GWAS (Lee, 2018) :

$$(\text{MLM}_0) : y_i = \beta_j X_{ij} + u_i + \varepsilon_i,$$

où ε_i désigne un terme d'erreur gaussien, i.i.d. et indépendant de X, et $(u_i)_{i=1,\dots,n}$ désigne l'effet aléatoire associé à l'individu i, issu du vecteur $u \sim \mathcal{N}_n(0, \sigma^2 K)$, avec $K \in \mathcal{M}_{n\times n}(\mathbb{R})$ la matrice d'apparentement tenant compte de la structuration de la population (Kang et al., 2008). Comme le calcul de la matrice d'apparentement et l'estimation du modèle est gourmand en ressources, il est fréquent de l'approcher en remplaçant l'effet aléatoire par des effets fixes constitués des premières composantes principales dérivées de l'analyse en composantes principales des génotypes standardisés (He et al., 2011). Dans ces modèles, le lien entre le SNP j et le phénotype est testé à l'aide d'un test de nullité du coefficient β_j.

Enfin, le modèle peut être estimé globalement (avec tous les individus de toutes les populations) ou bien décliné en R modèles indépendants, un pour chaque population. Nous noterons $(\text{MLM})_{0,r}$ le modèle pour la population $r \in \{1,\dots,R\}$.

Modèle hiérarchique bayésien (mashr). Récemment, la méthode `mashr` (Multivariate adaptive shrinkage, Urbut et al. (2019)) a été proposée pour estimer des effets communs et des effets propres à des conditions structurant une population d'individus. La méthode suppose que les effets du SNP j sur le phénotype dans les différentes populations, $\boldsymbol{\beta}_j = (\beta_{jr})_{r=1,\dots,R}$ se décompose selon une loi de mélange :

$$\mathbb{P}(\boldsymbol{\beta}_j | \boldsymbol{\pi}, \boldsymbol{U}) = \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_{k,l} \mathcal{N}_R(\boldsymbol{\beta}_j; 0, w_l U_k) \tag{1}$$

où $\mathcal{N}_R(\cdot; \boldsymbol{\mu}, \Sigma)$ désigne la densité d'une loi normale multivariée de dimension R de moyenne $\boldsymbol{\mu}$ et de matrice de variance-covariance Σ et $\pi_{k,l}$ sont les proportions du mélange. La méthode propose plusieurs patrons de matrices $U_k \in \mathcal{M}_{R\times R}(\mathbb{R})$, permettant de capturer les effets spécifiques ou communs des différentes sous-populations, et des w_l permettant de capturer des intensités d'effets variés.

Le modèle est estimé à l'aide d'une hypothèse de normalité pour les estimateurs $\hat{\boldsymbol{\beta}}_j$

$$\mathbb{P}(\hat{\boldsymbol{\beta}}_j|\boldsymbol{\beta}_j, V_j) = \mathcal{N}_R(\hat{\boldsymbol{\beta}}_j; \boldsymbol{\beta}_j, V_j) \tag{2}$$

qui permet, à partir de la distribution *a priori* (équation 1) et de la vraisemblance (déduite par l'équation 2), d'obtenir la distribution *a posteriori* :

$$\mathbb{P}(\hat{\boldsymbol{\beta}}_j|\boldsymbol{\pi}, \boldsymbol{U}, V_j) = \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_{kl}\mathcal{N}_R(\hat{\boldsymbol{\beta}}_j; 0, w_l U_k + V_j)$$

où V_j est la matrice de variances/covariances empirique entre races. Une fois le modèle estimé, les associations significatives sont obtenues par une approche *Local False Sign Rate* ($lfsr$) (Stephens, 2017) qui mesure la probabilité de commettre une erreur dans le signe de l'effet j déclaré significatif.

2.2 Cadre de simulation

Dans cette partie, nous détaillons la démarche de simulation mise en œuvre pour générer des données permettant l'évaluation des modèles et illustrée sur la figure 1.

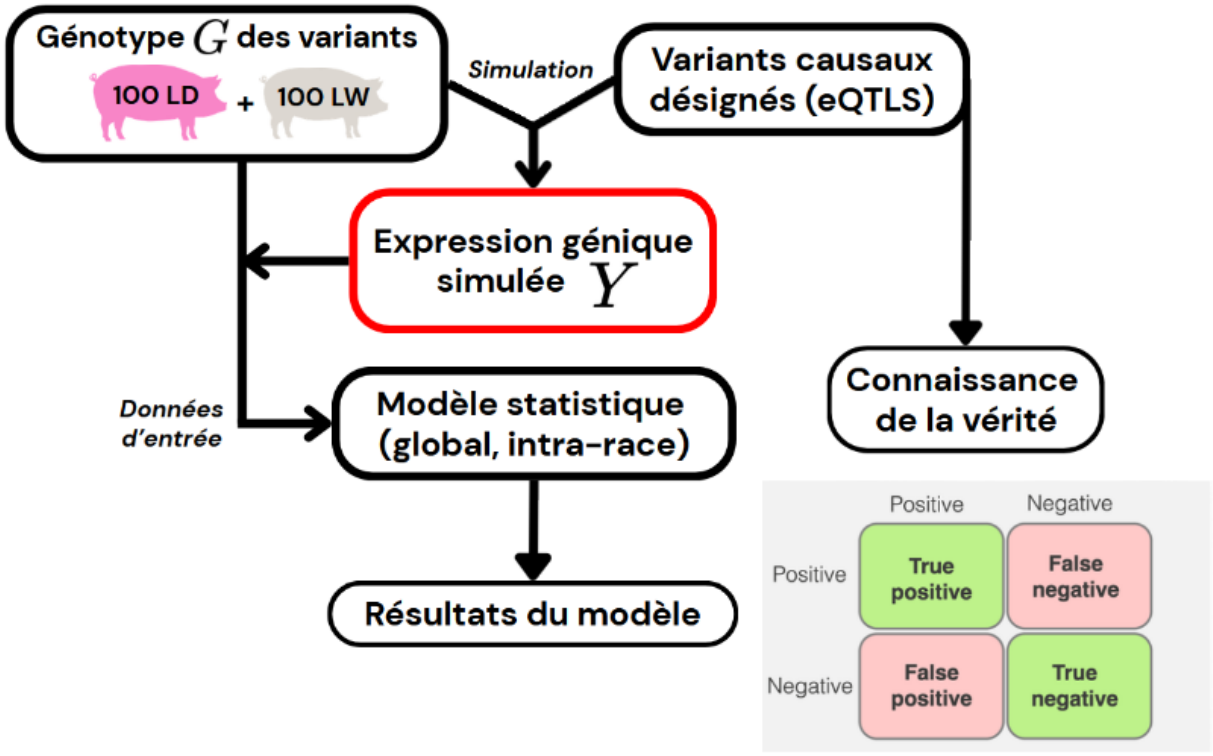

FIGURE 1 – Schéma simplifié de l'étude de simulation réalisée

Pour cette simulation, nous avons utilisé les génotypes réels de deux races porcines : Landrace (LD) et Large White (LW), issus de Crespo-Piazuelo et al. (2023). Cela nous a permis de maintenir une structure de dépendance réaliste entre les individus (associées à la parenté et à la stratification génétique) et entre les variants (déséquilibre de liaison).

Dans la suite, on appellera eQTL (*expression QTL*) tout variant ou SNP statistiquement associé à une variation de l'expression d'un gène. Pour chaque simulation, un échantillon de 5 000 variants a été choisi de manière aléatoire parmi ceux qui se trouvent sur le chromosome 18. En plus des variants, 100 gènes annotés dans la base de données `Ensembldb` et localisés sur le chromosome 18 ont été sélectionnés.

Pour chaque gène, 3 cis-eQTLs ont été choisis parmi les 5 000 SNPs, en sélectionnant au hasard des SNPs qui se trouvent dans une fenêtre de 100 000 paires de bases centrée autour de la séquence codante du gène. Ainsi, chaque simulation produit un ensemble de 300 paires {gène, eQTL}, constituant la « vérité » à retrouver.

2.2.1 Catégorisation des variants

Les variants ont d'abord été classés en trois catégories selon leur VAF ainsi que leurs fréquences spécifiques dans chacune des deux races considérées :

- Variants spécifiques à une race : la fréquence allélique du variant est strictement nulle dans l'une des deux races. Deux sous-catégories sont distinguées selon la race dans laquelle l'allèle est absent : `LD_specific` (absence dans la race LW) et `LW_specific` (absence dans la race LD),
- Variants contrastés entre races : deux sous-catégories sont définies, `LD_contrasted` et `LW_contrasted` selon la race avec la plus grande variabilité,
- Variants homogènes : ces variants sont regroupés dans la catégorie `homogeneous`.

2.2.2 Simulation des effets

La génération des effets repose sur un modèle de référence commun :

$$\beta_j^{(r)} = \text{Ber}\left(\{-1, 1\}; 0, 5\right) \times \mathcal{N}(1; 0, 1). \tag{3}$$

Pour les **eQTLs spécifiques** à une race, l'effet $\beta_j^{(r)}$ pour cette race est simulé selon le modèle de référence de l'équation (3). Plusieurs types d'effets biologiquement plausibles ont ensuite été générés pour les variants non spécifiques :

- **Effets identiques** : les deux races reçoivent le même effet, généré selon le modèle de référence de l'équation (3).
- **Effets différents** : les deux races reçoivent des effets de même signe mais d'intensités distinctes, tirées dans deux lois normales différentes : $\mathcal{N}(0.5, 0.1)$ et $\sim \mathcal{N}(1.5, 0.1)$. Le signe des deux effets, ainsi que leur attribution aux deux races respectives, sont déterminés aléatoirement selon la même loi de Bernoulli symétrique, de manière à ce qu'il n'y ait pas de direction privilégiée pour une race donnée.
- **Effets opposés** : les deux races reçoivent des effets de même intensité mais de signes inversés, simulés à partir de deux lois normales symétriques centrées sur des valeurs opposées $\mathcal{N}(1, 0.1)$ et $\mathcal{N}(-1, 0.1)$. L'attribution de ces deux effets aux races est également réalisée aléatoirement.

Les trois types d'effets ont été générés de manière équiprobable parmi les eQTL non spécifiques.

2.2.3 Simulation des données d'expression

L'expression simulée d'un gène pour un individu i dans une race r est obtenue à partir des effets simulés $\beta_j^{(r)}$ et perturbée par un bruit aléatoire contrôlant l'héritabilité du gène :

$$y_i^{(r)} = \sum_{j=1}^{m} \beta_j^{(r)} X_{ij}^{(r)} + \varepsilon_{ij}^{(r)}, \text{ avec } \varepsilon_i^{(r)} \sim \mathcal{N}(0, \sigma_r^2)$$

où σ_r^2, est calibrée de manière à atteindre un niveau d'héritabilité cible fixé à $h^2 = 0, 8$.

3 Premiers résultats et discussion

Pour l'instant, nous présentons uniquement les résultats du modèle linéaire mixte, l'évaluation de la méthode `mashr` étant toujours en cours.

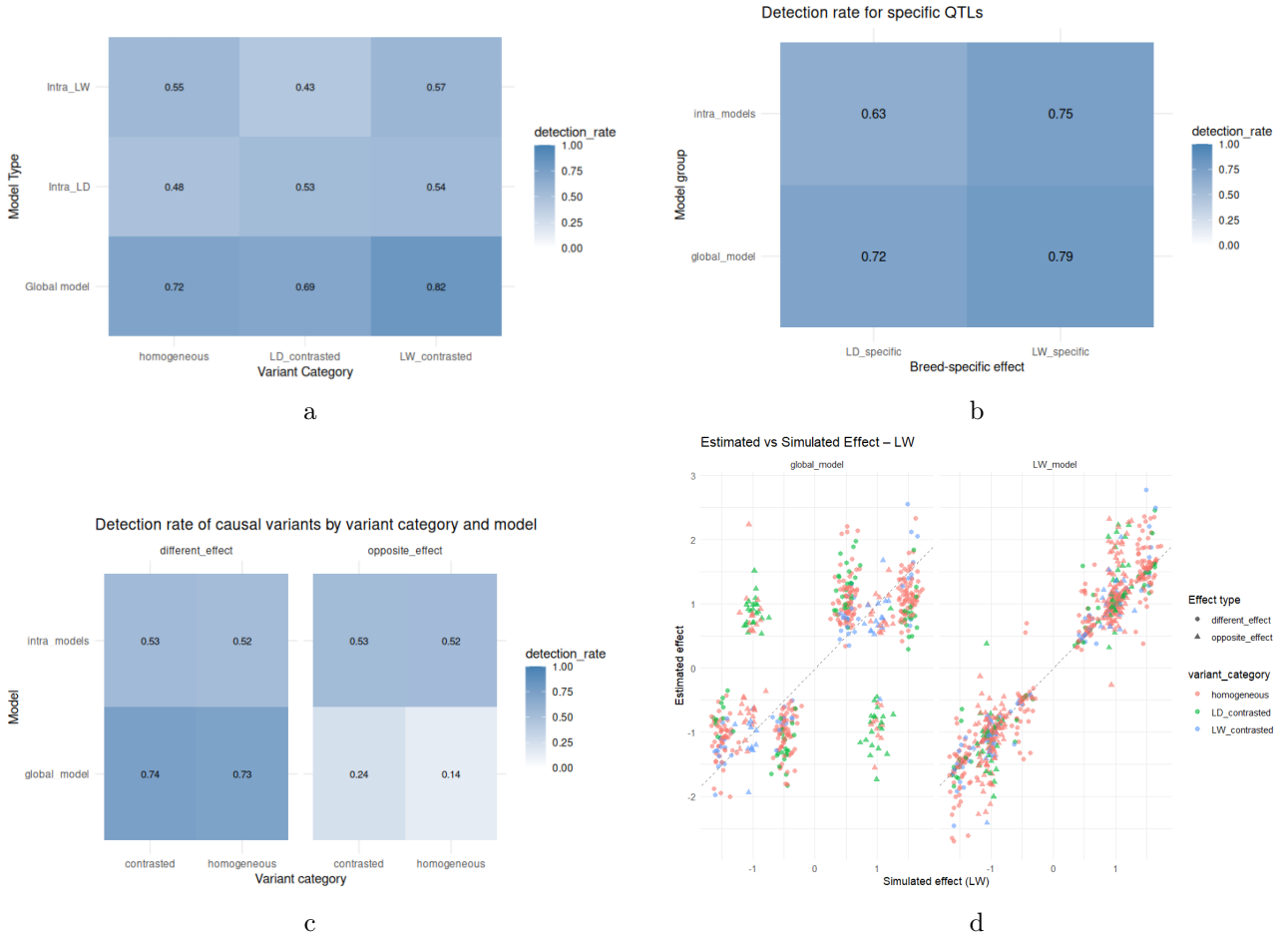

FIGURE 2 – a) c) carte de chaleur du taux de détection des eQTLs à effets identiques ou complexes en fonction des catégories de fréquences, b) carte de chaleur du taux de détection des eQTLs spécifiques à la race LD ou LW, d) comparaison des β estimés par les modèles faces aux β complexes simulés en amont.

Le modèle global identifie en général plus d'eQTLs que les modèles spécifiques aux races, indépendamment de la catégorie de la VAF (figure 2a). Les modèles intra-races se révèlent plus performants pour détecter les eQTLs contrastés en faveur de leur race associée. Le modèle global s'avère légèrement plus performant que les modèles intra-races (figure 2b). La figure 2c montre que le modèle global détecte efficacement les variants présentant des effets différents, avec une performance comparable à celle obtenue pour les variants ayant des effets identiques. En revanche, sa capacité à détecter les variants à effets opposés entre les races s'effondre. La figure 2d montre l'estimation des modèles en fonction des types d'effets complexes simulés. Pour les variants à effets différents, deux groupes distincts sont observés, conformément aux distributions simulées. Le modèle intra-race reproduit correctement cette structure, tandis que le modèle global présente des résultats plus hétérogènes. Les erreurs d'estimation sont surtout observées pour les variants variables chez Landrace, alors que les variants `LW_contrasted` sont mieux estimés ; les effets opposés demeurent les plus difficiles à prédire.

4 Conclusion et perspectives

Les résultats obtenus suggèrent que les modèles globaux présentent une meilleure performance lorsqu'il s'agit de détecter des variants à effet commun entre races, grâce à leur meilleure puissance statistique. En revanche, leur capacité à détecter des eQTLs et à estimer correctement leurs effets en présence d'effets opposés entre populations est fortement compromise. Les modèles intra-races offrent une alternative intéressante pour explorer les effets spécifiques et différents entre races, mais au prix d'une perte notable de puissance.

Les modèles hiérarchiques bayésiens (`mashr`) apparaissent comme des pistes prometteuses pour dépasser ces limitations, en combinant robustesse et flexibilité dans la prise en compte de l'hétérogénéité inter-populations. Les prochains mois permettront d'évaluer leur efficacité dans un contexte multi-population où R est généralement petit contrairement au cadre initial de la méthode.

Remerciements

La thèse de Kossi Julien Kowou est financée conjointement par l'Agence Nationale de Recherche dans le cadre du projet ANR-22-PEAE-4 (France 2030) et par le métaprogramme INRAE DIGIT-BIO.

Références

D. Crespo-Piazuelo, H. Acloque, O. González-Rodríguez, M. Mongellaz, M.-J. Mercat, M. C. Bink, A. E. Huisman, Y. Ramayo-Caldas, J. P. Sánchez, and M. Ballester. Identification of transcriptional regulatory variants in pig duodenum, liver, and muscle tissues. *GigaScience*, 12 :giad042, 2023.

H. He, X. Zhang, L. Ding, T. M. Baye, B. G. Kurowski, and L. J. Martin. Effect of population stratification analysis on false-positive rates for common and rare variants. In *BMC proceedings*, volume 5, page S116. Springer, 2011.

H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3) :1709–1723, 2008.

J. H. Ko, A. Rau, and N. Vialaneix. Analyse de la spécificité des associations génétiques dans les études multi-population. In *Journées de Statistique de la SFdS*, 2024.

C. Lee. Genome-wide expression quantitative trait loci analysis using mixed models. *Frontiers in Genetics*, 9 :341, 2018.

M. Stephens. False discovery rates : a new deal. *Biostatistics*, 18(2) :275–294, 2017.

S. M. Urbut, G. Wang, P. Carbonetto, and M. Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature genetics*, 51(1) :187–195, 2019.

I. van den Berg and I. M. MacLeod. The impact of qtl sharing and properties on multi-breed gwas in cattle : a simulation study. *Animal Production Science*, 63(11) :996–1007, 2023.