Analyse de la spécificité des associations génétiques dans les études multi-population.

Jeong Hwan Ko[1,2], 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*
{jeong-hwan.ko,andrea.rau,nathalie.vialaneix}@inrae.fr

Résumé. Les études d'association génomique (GWAS) sont un outil essentiel pour comprendre la relation entre les variants génétiques et les phénotypes complexes dans les populations humaines ainsi que dans les espèces agricoles. La structuration de population pour les animaux d'élevage est la conséquence des forces évolutives mais également d'une forte sélection anthropique, conduisant à des races hautement spécialisées répondant à divers besoins. Cette forte structuration génétique doit être correctement prise en compte dans un GWAS pour éviter la détection de faux positifs. À ce jour, peu d'études ont évalué l'impact conjoint d'une forte structuration multi-population et du choix du modèle sur les résultats d'un GWAS. Pour combler ce manque, nous considérons le cas d'un GWAS dans trois races porcines pour l'expression d'un seul gène, HUS1, en utilisant les données de séquençage en génome entier du chromosome 18. Nous avons spécifiquement évalué l'impact du choix d'un modèle linéaire en utilisant un modèle avec ou sans un effet fixe de race et/ou effet aléatoire lié à l'apparentement génétique des animaux. Nous étudions alors le nombre et la similarité des associations détectées. Également, nous avons étudié l'impact des fréquences alléliques spécifiques à une race sur des associations génétiques globales ou spécifiques à une race. Comme attendu dans ce contexte, nos résultats privilégient l'utilisation du modèle linéaire mixte classique qui utilise l'apparentement comme effet aléatoire. Nos résultats soulignent également la nécessité de développer de nouveaux modèles mieux capables de distinguer les associations génétiques spécifiques à une population de celles partagées par plusieurs populations.

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

Abstract. Genome-Wide Association Studies (GWAS) are an essential tool for understanding the relationship between genetic variants and complex phenotypes in human populations as well as livestock and crop species. Population structure in livestock is shaped not only by evolutionary forces but also by strong human-mediated selection, leading to highly specialized breeds that meet various needs. This strong genetic structure must be correctly accounted for within a GWAS to avoid the detection of false positives. To date few studies have evaluated the joint impact of strong population structure and GWAS model choice on downstream results. To address this gap, we consider the case of a GWAS in three pig breeds for the expression of a single gene, HUS1, using whole genome sequencing data from chromosome 18. We specifically evaluated the impact of linear model choice using a fixed and/or random breed effect to adjust for breed structure on the number and similarity of detected associations. We notably investigated the impact of breed-specific minor allele frequencies on

significant global or breed-specific genetic associations. As expected, our results favor the use of the classic linear mixed model using kinship as a random effect in this setting. Our results also highlight the need to develop new models better able to distinguish population-specific from globally shared genetic associations.

Keywords. multi-population, genetic association, linear mixed model

1 Introduction

Les études d'associations génétique à grande échelle ou Genome-Wide Association Studies (GWAS) sont utilisées pour étudier l'association entre variants génétiques dans le génome et un phénotype d'intérêt. À la différence des populations humaines, les espèces agricoles ont généralement été soumises à une sélection très forte avec des objectifs variés, menant ainsi à une structuration en races pour des animaux d'élevage ou en écotypes pour les espèces végétales, que l'on désignera dans la suite sous le terme générique de « populations ». Si l'effet des populations est parfois pris en compte dans les modèles d'association (par l'ajout d'un effet fixe ou d'un effet aléatoire par exemple), très peu d'études examinent l'impact conjoint de cette structuration et du choix du modèle sur les résultats des analyses.

Une première étude de simulation a été réalisée par I. van den Berg *et al.* [5] dans laquelle 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 trait phénotypique) ont été utilisés dans trois cadres différents : un premier où la fréquence de l'allèle mineur (MAF) était similaire dans les trois races bovines étudiées, un second où la MAF était variable dans les trois races, et un troisième où les QTL étaient spécifiques aux races (ce qui signifie que la MAF au sein de cette race est non nulle pour cette race et nulle pour les autres). Les auteurs ont regardé l'impact de ces trois scénarios à l'aide de plusieurs modèles GWAS dont un modèle intra-races (réalise un GWAS au sein d'une seule race), un modèle global réalisant un GWAS en prenant toutes les races confondues, et un modèle intégrant une matrice d'apparentement afin de prendre en compte les structures de population. Les résultats montrent (1) un manque de puissance des analyses intra-race ou globalement des analyses portants sur les variants dont la MAF est faible, et (2) un impact important du type de modèle sur les résultats, même lorsque l'association entre le variant et le phénotype est uniforme pour toutes les races. Cependant, cette étude ne suffit pas, d'une part car elle se base uniquement sur des données simulées et d'autre part, parce qu'elle ne considère pas le cas où l'association elle-même peut être spécifique d'une population.

Notre objectif ici est donc d'explorer cette question de manière plus approfondie, afin de proposer une évaluation exhaustive de l'impact conjoint du choix de modèle et des variations de MAF ou d'association inter-populations sur les résultats. Dans cette communication, nous présentons une étude réalisée sur des données réelles porcines multi-races. Le but est d'étudier la différence de résultats entre des études d'association globales (toutes les races confondues, avec ou sans prise en compte d'un effet race) et des études où les modèles sont estimés séparément pour chaque race.

La suite du résumé présente les données et divers modèles que nous avons mis en œuvre et les données sur lesquelles nous avons réalisé cette étude (section 2) puis commente les premiers résultats obtenus (section 3). Des résultats plus approfondis seront discutés durant la conférence.

2 Une étude comparative des différentes approches pour l'analyse GWAS multi-population

2.1 Modèles étudiés

Dans toute cette partie, on décrit le cadre formel d'une étude GWAS multi-population avec n observations. Dans cette étude, on cherche à expliquer les observations d'un phénotype $\mathbf{y} = (y_i)_{i=1,\dots,n}$, que l'on supposera numérique, par les observations de p SNPs $\mathbf{X} = (X_{ij})_{i=1,\dots,n,\, j=1,\dots,p}$ tels que $X_{ij} \in \{0,1,2\}$ correspondant au nombre d'allèles alternatifs porté par l'individu i au locus p. En outre, les n individus sont structurés en K sous-populations (des races ou des écotypes par exemple), représentées par la donnée du vecteur $\mathbf{z} = (z_i)_{i=1,\dots,n}$ tel que $z_i \in \{1,\dots,K\}$. L'objet de cette section est de décrire les différents modèles, prenant en compte ou non l'information de populations, que nous avons étudiés.

Modèle linéaire simple. Le modèle le plus simple pour les analyses GWAS est le modèle linéaire simple expliquant $\mathbf{y}$ à partir de $X_{\cdot j}$ pour $j = 1,\dots,p$. Si l'effet de la population est ignorée, cette approche peut se résumer à l'estimation de p modèles linéaires

$$(\mathrm{ML}_0): \qquad y_i = \beta_j X_{ij} + \epsilon_i,$$

avec ϵ_i les termes d'erreur gaussien, i.i.d. et indépendants de X. Ce modèle conduit à l'estimation d'une p-valeur pour chaque test $\mathrm{H}_0 : \ll \beta_j = 0 \gg$ qui correspond au test de l'association du SNP j avec le phénotype.

Ce modèle peut être amélioré pour prendre en compte un effet fixe additif correspondant à la population. On obtient alors

$$(\mathrm{ML}_1): \qquad y_i = \beta_j X_{ij} + \alpha_k \mathbf{1}_{\{z_i=k\}} + \epsilon_i,$$

où α_k est l'effet de la population k de l'individu i.

Enfin, le modèle (ML_0) peut être décliné en K modèles indépendants, un pour chaque population :

$$(\mathrm{ML}_{0,k}): \qquad y_i = \beta_j X_{ij} + \epsilon_i, \qquad \forall\, i:\ z_i = k.$$

Dans ce dernier cas, l'étude de l'association d'un SNP j avec le phénotype conduit à l'obtention de K p-valeurs, une pour chaque population.

Modèle mixte. Toutefois, il est généralement naïf de considérer que les effets de population peuvent être omis ou modélisés au travers d'un simple effet fixe additif. Plusieurs travaux [4, 3, 6] proposent l'utilisation d'un modèle mixte avec une matrice permettant de modéliser la proximité génétique entre individus : la matrice d'apparentement [4].

De manière plus précise, ces travaux considèrent que l'effet aléatoire, u_i, du patrimoine génétique global d'un individu i peut être modélisé comme une somme de petits effets aléatoires :

$$u_i = \sum_{h\in\{0,1,2\}} L_{i,h}^\top v_h$$

où $L_{i,h}$ est le vecteur de taille p contenant le codage disjonctif de la valeur des p SNP de l'individu i, $X_{i.}$, pour la valeur $h \in \{0,1,2\}$ et v_h est un vecteur aléatoire de taille p dont les coordonnées, $v_{j,h}$ sont des variables aléatoires i.i.d. de loi $\mathcal{N}(0,\sigma^2)$. En notant L_h la matrice $(n \times p)$ $[L_{1,h},\ldots,L_{n,h}]^\top$, on montre alors que

$$\mathrm{Var}(u) = \sigma^2 \sum_{h\in\{0,1,2\}} L_h^\top L_h.$$

Pour la suite, nous noterons la matrice $Q = \sum_{h\in\{0,1,2\}} L_h^\top L_h$.

Les modèles mixtes d'association incluant l'information d'apparentement s'écrivent ainsi :

$$(\mathrm{MLM}_0): \qquad y_i = \beta^\top X_{i.} + u_i + \epsilon_i,$$

avec $X_{i.}$ le vecteur des p valeurs des SNP de l'individu i, et $(u_i)_{i=1,\ldots,n} \sim \mathcal{N}_n(0,\sigma^2 Q)$ comme décrit précédemment (dans le cas de ce modèle, les colonnes de $\mathbf{X}$ sont préalablement centrées et réduites). Dans ces modèles, l'association entre le SNP j et le phénotype est testée par un test de la nullité du coefficient β_j du vecteur β.

Les mêmes variantes de ce modèle que celles proposées pour le modèle linéaire simple sont également considérées, soit un modèle incluant l'effet population en effet fixe :

$$(\mathrm{MLM}_1): \qquad y_i = \beta^\top X_{i.} + u_i + \alpha_k \mathbf{1}_{\{z_i=k\}} + \epsilon_i,$$

et la déclinaison du modèle (ML_0) dans chacune des populations :

$$(\mathrm{MLM}_{0,k}): \qquad y_i = \beta^\top X_{i.} + u_i + \epsilon_i, \qquad \forall\, i:\ z_i = k.$$

Chacun de ces modèles donne une p-valeur par SNP (soit p p-valeurs par modèle).

2.2 Données

Ces différents modèles ont été comparés sur les données décrites dans [1]. De manière plus précise, nous avions à notre disposition des données appariées de séquençage du génome entier (WGS) et transcriptomique (expression des gènes) du duodenum sur $n = 300$ cochons issus de trois races, dont deux races commerciales (Large White et Landrace) et une race considéréé comme plus robuste (Duroc). Les effectifs des trois races sont équilibrés (100

individus par race). Une analyse en composantes principales (ACP) des génotypes (figure 1 gauche) sur l'ensemble des SNP de tous les chromosomes montre une séparation génétique nette des trois races.

Par soucis de simplicité, les premières analyses ont été limitées à l'étude des associations pour les variants du chromosome 18 uniquement ($p = 665\,148$ variants) sur un seul phénotype correspondant à l'expression du gène HUS1. Ce gène était décrit dans [1] comme étant régulé génétiquement (càd comme ayant des associations significatives avec des SNP) dans les trois tissus étudiés dans l'article (duodenum, muscle et foie) et il est connu pour réguler la quantité de matière grasse intramusculaire.

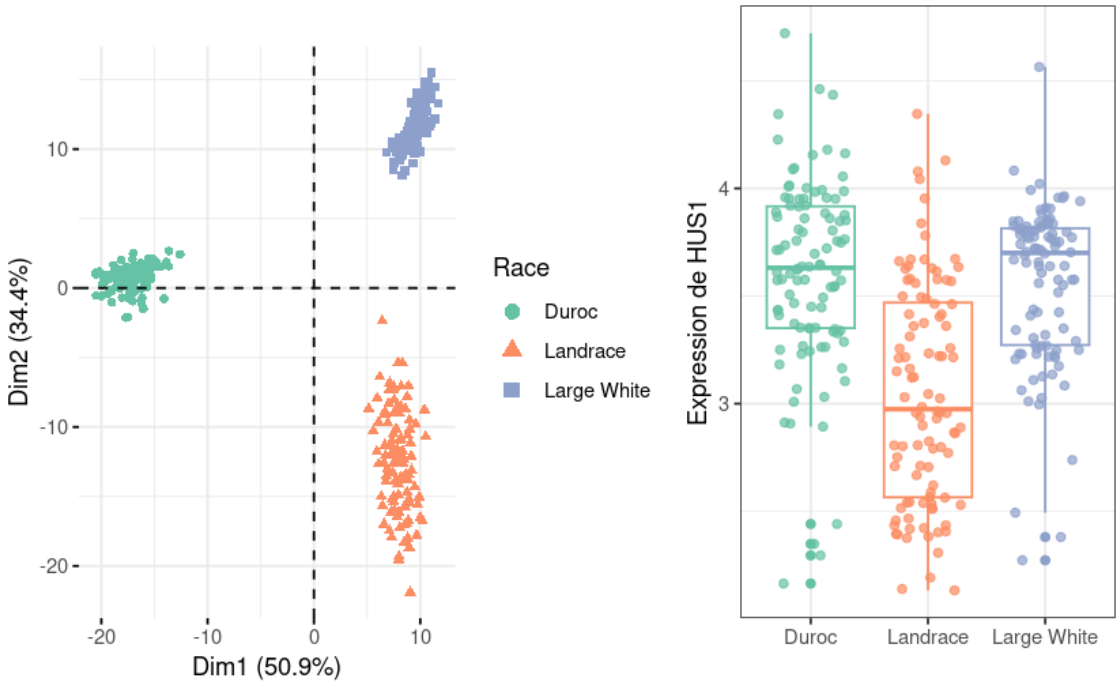

FIGURE 1 – Gauche : ACP sur les génotypes des individus (la matrice $\mathbf{X}$, utilisée pour l'ACP, a été préalablement centrée et réduite). Droite : Distribution de l'expression du gène HUS1 dans le duodenum selon la race ($n = 300$ individus).

La figure 1, droite illustre la distribution de l'expression du gène HUS1 selon la race. On observe que, selon les races porcines, le niveau d'expression du gène varie, ce qui peut s'expliquer soit par une expression différentielle de ce gène selon la race mais aussi par une régulation différente des SNP associés à ce gène selon la race (que celle-ci provienne d'une différence dans la fréquence allélique de ces SNP ou d'une différence dans leur niveau d'association avec l'expression du gène).

3 Premiers résultats et discussion

Les analyses ont été effectuées avec le logiciel R : la fonction `lm` a été utilisée pour l'estimation des modèles linéaires et la fonction `GWAS` du package **rrBLUP** [2] pour l'estimation des modèles linéaires mixtes (ce package contient aussi la fonction `A.mat` permettant de calculer la matrice d'apparentement).

Pour chaque modèle testé, les p-valeurs de l'association de l'ensemble des variants avec l'expression du gène considéré sont obtenues et les p-valeurs sont corrigées pour le contrôle des tests multiples (contrôle du FWER, correction de Bonferroni [**?**]). Une association est déclarée significative si la p-valeur ajustée est inférieure à un seuil de $\alpha = 5 \times 10^{-8}$.

3.1 Impact des différents modèles sur le nombre d'associations significatives détectées

La figure 2 montre le nombre d'associations détectées par chaque modèle. Dans presque tous les modèles (à l'exception des modèles spécifiques à la race Landrace), on observe une inflation très importante du nombre d'associations positives pour le modèle linéaire simple. Ceci est un résultat attendu et s'explique par la forte dépendance des génotypes des animaux, notamment au sein d'une même race. Or, le modèle linéaire considère les animaux comme i.i.d., ce qui conduit à une inflation du nombre d'associations (faussement) positives détectées par le modèle.

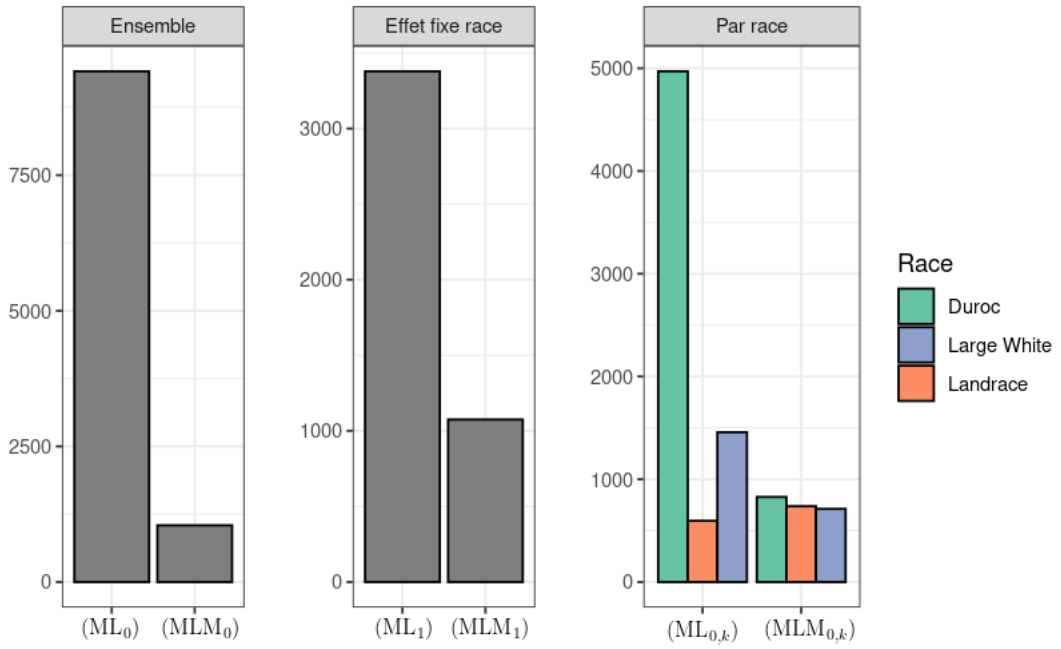

FIGURE 2 – Nombre d'associations détectées par méthode pour un contrôle du FWER au niveau $\alpha = 5 \times 10^{-8}$.

3.2 Associations communes ou spécifiques aux différents modèles

La figure 3 illustre les associations détectées de manière commune ou spécifique par les différentes approches. Notons que nous n'avons pas inclus les modèles $(\mathrm{ML}_{0,k})$ et $(\mathrm{MLM}_{0,k})$ dans ce graphique par soucis de clarté.

De manière intéressante, pour le modèle linéaire mixte estimé sur l'ensemble des individus, la présence ou l'absence d'un effet fixe race au modèle ne modifie que très marginalement les résultats obtenus : les deux modèles détectent, en effet, 1043 associations communes, sur 1046 et 1075 associations détectées au total par chacun des deux modèles respectivement.

À l'inverse, pour les modèles linéaires simples, l'ajout de l'effet fixe race a un impact considérable sur les résultats : on trouve 1026 associations communes, sur 9414 et 3378 associations détectées au total par chacun des deux modèles respectivement. On constate également que la quasi totalité (1042 sur 1046) des associations trouvées par le modèle linéaire mixte sont également retrouvées par le modèle linéaire simple avec effet race. L'effet fixe race semble donc contenir, à lui seul, une partie importante de l'information sur la non indépendance des individus (ce qui était attendu compte tenu de l'ACP de la figure 1 gauche).

Dans une moindre mesure, les associations détectées par le modèle linéaire mixte sont aussi majoritairement détectées par le modèle linéaire simple sans effet race (850 sur 1075).

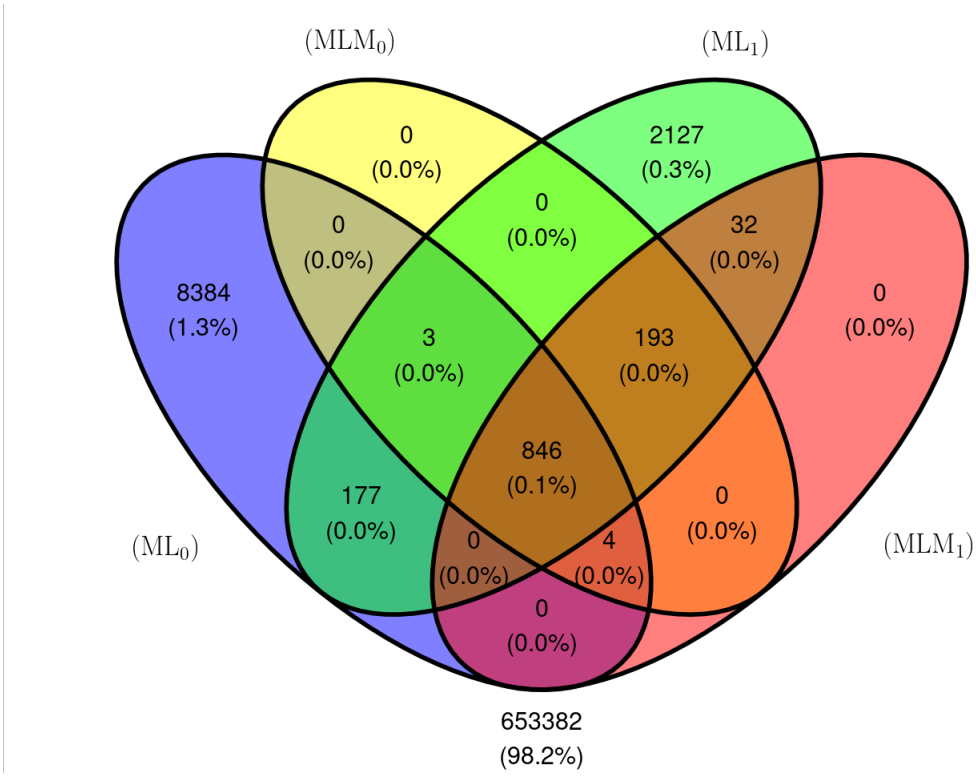

FIGURE 3 – Associations significatives communes et spécifiques des différents modèles.

3.3 Impact des différences entre fréquences des allèles mineurs sur des associations spécifiques par race

Enfin, nous abordons ici un point qui sera au cœur des discussions de notre présentation. Les différences génétiques entre les trois races de l'étude sont importantes et certains des variants sont eux-mêmes spécifiques à une race donnée ou présentent une variation de la fréquence de l'allèle mineur entre les différentes races. Nous illustrons ce propos avec la figure 4 : dans celle-ci, le panneau de gauche présente un variant observé dans les trois races (avec des fréquences alléliques mineures très différentes entre les trois races) mais dont l'association avec l'expression de HUS1 est détectée comme significative uniquement dans la race Duroc (et négative uniquement dans cette race aussi). Dans le panneau de droite, une deuxième situation est présentée : le variant est détecté comme étant significativement associé à la race Duroc, mais pas aux autres races (le variant est même fixé dans les autres races). Dans les deux cas, ces variants n'ont pas d'association significative avec l'expression du gène HUS1 dans le modèle linéaire mixte (MLM_0) : ces deux exemples simples illustrent donc l'impact très important de la bonne prise en compte de l'effet population dans les modèles.

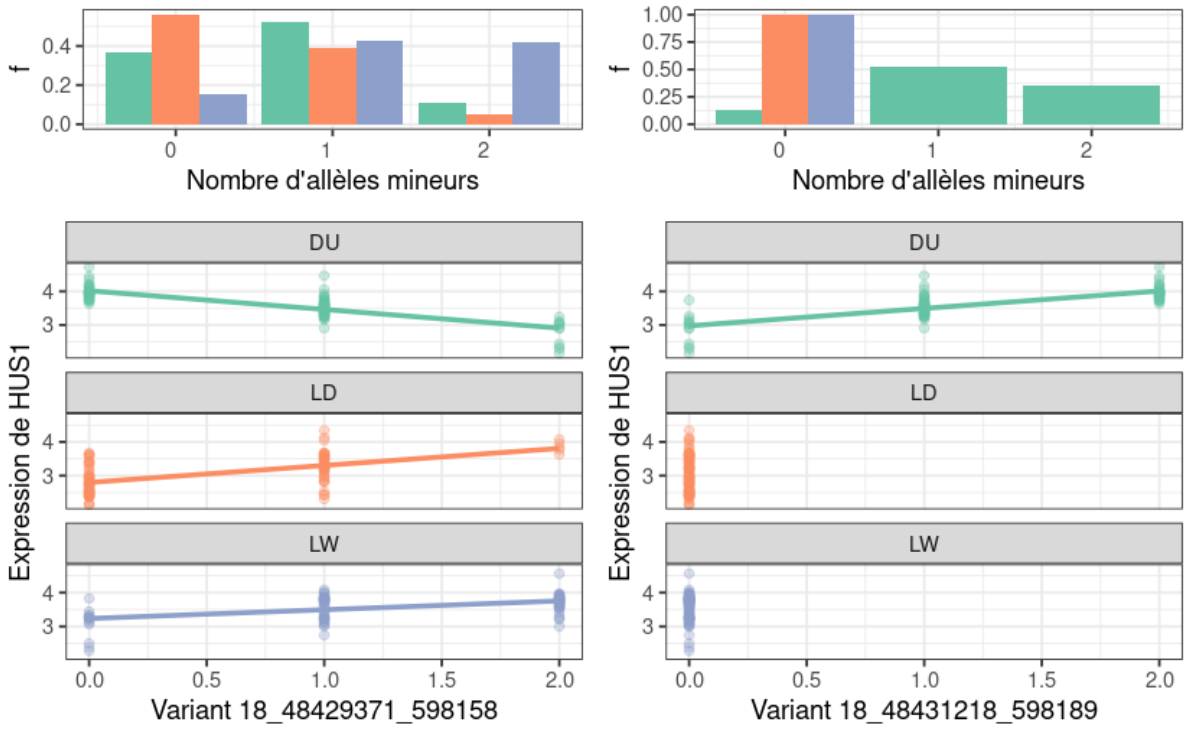

FIGURE 4 – Fréquence allélique et expression du gène HUS1 en fonction de l'allèle mineur au sein des races. L'identifiant du SNP trouvé en appliquant le modèle mixte au sein d'une race.

4 Conclusion

Une analyse exhaustive de l'impact conjoint des modèles et des différences de fréquences alléliques sur les résultats des tests d'association est proposée. Cette analyse souligne la disparité en termes de qualité d'analyse entre les modèles linéaires simples et les modèles linéaires mixtes dans un contexte d'étude multi-population en GWAS, ici en utilisant des données porcines réelles. Concernant les modèles linéaires où chaque variant génétique est considéré comme un prédicteur indépendant du trait phénotypique sans tenir compte de la structure de population, il a été constaté qu'une telle approche entraîne un taux de faux positifs plus élevé dans un contexte de structuration forte en populations non-indépendantes.

Ainsi, une approche plus sophistiquée, telle que les modèles linéaires mixtes intégrant une matrice d'apparentement, s'avère importante pour réaliser des tests d'associations plus précis dans ce contexte. Cette étude souligne donc la nécessité de développer un modèle plus adapté capable de distinguer les associations spécifiques aux populations des associations communes à plusieurs populations, notamment dans le cas des différences marquées de fréquences alléliques.

Remerciements

Ce travail a bénéficié d'une aide d'État gérée par l'Agence Nationale de Recherche au titre de France2030 portant la référence « ANR-22-PEAE-4 ».

Bibliographie

[1] Daniel Crespo-Piazuelo, Hervé Acloque, Olga González-Rodríguez, Mayrone Mongellaz, Marie-José Mercat, Marco C A M Bink, Abe E Huisman, Yuliaxis Ramayo-Caldas, Juan Pablo Sánchez, and Maria Ballester. Identification of transcriptional regulatory variants in pig duodenum, liver, and muscle tissues. *GigaScience*, 12:giad042, 2023.

[2] Jeffrey B. Endelman. Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, 4(3):250–255, 2011.

[3] Hyun Min Kang, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-yee Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42:348–354, 2010.

[4] Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[5] Irene van den Berg and Iona M. MacLeod. The impact of QTL sharing and properties on multi-breed GWAS in cattle: a simulation study. *Animal Production Science*, 63(10-11):996–1007, 2023.

[6] Zhang Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J. Todhunter, Hemant K. Tiwari, Michael A. Gore, Peter J. Bradbury, Jianming Yu, Donna K. Arnett, Jose M. Ordovas, and Edward S. Buckler. Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 42:355–360, 2010.